SCIENCE IN PROGRESS
SIXTEENTH SERIES

SIXTEENTH SERIES

Edited by WALLACE R. BRODE

Science in Progress

BY

HUGH TAYLOR
LYLE B BORST
RENÉ DUBOS
NORMAN H CROMWELL
ALPHONSE CHAPANIS
ERNEST C POLLARD
C. J PHILLIPS
J BRONOWSKI
G LEDYARD STEBBINS
C M SLIEPCEVICH
TALBOT H WATERMAN
W H PICKERING

Set in Baskerville type by

The McKay Press, Midland, Michigan

and printed in the United States of America by

The Carl Purington Rollins Printing-Office of

the Yale University Press, New Haven, Connecticut.

Distributed in Canada by McGill University Press.

Library of Congress catalog card number: 39-14778

CONTENTS

Preface, WALLACE R. BRODE ix

1 Academia and Industry—Their Mutual Influence 1
HUGH TAYLOR

2. Liquid Helium 21
LYLE B. BORST

3. Humanistic Biology 45
RENÉ DUBOS

4 Chemical Carcinogens, Carcinogenesis, and Carcinostasis 69
NORMAN H. CROMWELL

5. Color Names for Color Space 105
ALPHONSE CHAPANIS

6. The Fine Structure of the Bacterial Cell and the Possibility of Its Artificial Synthesis 133
ERNEST C. POLLARD

7. The Strength and Weakness of Brittle Materials 171
C. J PHILLIPS

8 The Logic of the Mind 217
J BRONOWSKI

9. From Gene to Character in Higher Plants 239
G. LEDYARD STEBBINS

10 Liquefied Natural Gas—A New Source of Energy 273
C M SLIEPCEVICH

11. Systems Analysis and the Visual Orientation of Animals 323
TALBOT H WATERMAN

12 The Selection of Space Experiments 373
W. H PICKERING

Index 383

PREFACE

The first of the series of *Science in Progress* appeared in 1939 and contained the Sigma Xi National Lectures of a two year-period (1937-38) In this present volume, the sixteenth in the series, are the Sigma Xi—RESA National Lectures for the academic year 1964-65. The increase in popularity and demand for the Sigma Xi National Lectures has resulted in a marked increase in the number of lectures each year The earlier plan had required that each lecturer supply a prepared manuscript of his lecture to the editor of *American Scientist* After publication in that journal, the manuscripts were collated to produce the volumes known as *Science in Progress*

The first to ninth volumes, covering the Sigma Xi National Lectures from 1937 to 1954, were edited by George A Baitsell The tenth and eleventh volumes of the *Science in Progress* series, covering the Sigma Xi—RESA National Lectures from 1955-56 and 1957-58, were edited by Hugh Taylor, also the editor of *American Scientist* Commencing with the twelfth volume of the series and including the present one, the increased number of Sigma Xi National Lectures has provided sufficient material to form a complete volume in one year

The great demand for Sigma Xi lectures has revealed the fact that many possible lecturers, having published portions of their material in recent journals, were unwilling to present lectures on the subject if they had to submit a manuscript Others preferred to publish in specialized journals, and still others did not wish to prepare a definitive manuscript at the time In order to find enough lecturers to supply

our needs, the Society decided in 1965-66 not to require a completed manuscript as a condition of the lectureship but to make it optional on the part of each lecturer. With some fifteen to twenty lecturers in a season, it would be quite difficult to collect and publish the entire series, and hence it has been decided by the Executive Committee of Sigma Xi to terminate publication of the lecture series with this volume of *Science in Progress.*

BIOGRAPHICAL DATA

LYLE B. BORST, Ph.D., Chicago (1941). Professor of Physics at the State University of New York at Buffalo. A nuclear physicist with wide experience in research and direction of programs at the University of Chicago, Oak Ridge, and Brookhaven. His researches on liquid helium originate from neutron studies of molecular crystals at liquid helium temperatures. *(Pacific Tour)*

J. BRONOWSKI, Ph.D., Cambridge (1933). A distinguished mathematician who is now serving as a Senior Fellow in the Salk Institute for Biological Studies in San Diego, California. He has written and lectured on modern theorems in mathematical logic and their limitations. His writings often approach and enter into the common areas of science and the humanities. The Phi Beta Kappa—Sigma Xi lecturer at the meeting of the American Association for the Advancement of Science in Berkeley, California, in December 1965.

ALPHONSE CHAPANIS, Ph.D., Yale (1963). Professor of Psychology and Industrial Engineering at The John Hopkins University. He has been recognized for his leading contributions to the field of engineering psychology and the introduction of human factors into our industrial developments. *(Mid-West Tour)*

NORMAN H. CROMWELL, Ph.D., Minnesota (1939) Regent's Professor and Chairman of the Department of Chemistry at the University of Nebraska. An authority on carcinogenesis and cancer therapy, and a consultant to the pharmaceutical industry and the United States Public Health Service in this area of chemical research *(Southwest Tour)*

RENÉ DUBOS, Ph.D., Rutgers (1927). Professor in The Rockefeller University. For a period of forty years he has served his university and the science of microbiology in the development of new solutions to the pressing problems of the experimental pathologists As an editor and author he has created a humanistic area of science which has added considerably to his international recognition. He was the Phi Beta Kappa—Sigma Xi lecturer at the Montreal meeting of the American Association for the Advancement of Science in December 1964.

C J. PHILLIPS, M A, Oberlin (1931) Professor of Ceramics at Rutgers, the State University of New Jersey. A leading author and consultant in the field of glass technology, with the experience of direct research and consulting service with Corning Glass Works, the Pittsburgh Plate Glass Company, and the Dunbar and Indiana Glass Companies. *(Southeast Tour)*

W. H. PICKERING, Ph D, California Institute of Technology (1936) Since 1954 he has served as the Director of the Jet Propulsion Laboratory of the California Institute of Technology He has directed the *Explorer, Ranger,* and *Mariner* spacecraft programs which have provided much of our advancing information on the moon, Mars, and interplanetary space He was the 1965 recipient of the Procter Prize of the Scientific Research Society of America

ERNEST C POLLARD, Ph D, Cambridge (1932) Professor of Biophysics and Chairman of the Department of Biophysics at Yale University until 1961 Since 1961 he has created and currently heads a Department of Biophysics at Penn-

sylvania State University. As a pioneer in a new area involving the biological and physical sciences he has initiated new research programs in the effects of heat and radiant energy on enzymes, viruses, and other organic forms. *(Northeast Tour)*

SLIEPCEVICH, Ph.D., University of Michigan (1948). Professor and Chairman of the Department of Chemical Engineering at the University of Oklahoma. He is currently serving as the Chairman of the School of General Engineering at the University of Oklahoma. An authority on liquefied natural gas and a consultant to industrial users and suppliers in this important industrial field. (Northeast Tour)

G. LEDYARD STEBBINS, Ph.D., Harvard (1931). Professor of Genetics at the University of California (Davis). An authority on organic evolution and organizer of the Department of Genetics at his university. His book on *Variation and Evolution in Plants* is recognized as a leading work in that field. *(Pacific Tour)*

HUGH TAYLOR, D. Sc., Liverpool (1914). For over half a century Dean Taylor has served this nation and in particular Princeton University in the area of chemistry. He was Chairman of the Princeton Department of Chemistry and Dean of the Graduate School. Since his emeritus status in 1958 he has been president of the Woodrow Wilson National Fellowship Foundation. His native country, Great Britain, with which he has held close ties, has conferred on him the honor of naming him Knight Commander of the Order of the British Empire, and in this country he has also been highly honored and recognized. He has served the Society of the Sigma Xi and RESA as editor of *American Scientist* for more than a decade and he was the editor of two volumes of *Science in Progress*. He received the Procter Prize of the Scientific Research Society of America (RESA) in 1964.

TALBOT H. WATERMAN, Ph.D., Harvard (1943). Professor of Biology in the Josiah Willard Gibbs Research Laboratories of Yale University. An authority in a multidisciplinary approach to systems analysis and computer techniques for dealing with biological phenomena. *(Mid-Atlantic Tour)*

This present volume in the *Science in Progress* series presents lectures given in 1965 and 1966, including the Phi Beta Kappa–Sigma Xi lectures at the meetings in Berkeley and Montreal of the American Association for the Advancement of Science in December 1964 and 1965, and the RESA–Procter Prize lectures by Hugh Taylor and W. H. Pickering.

The editor of this volume wishes to express his appreciation to the separate authors of the lectures who have cooperated in the revision and publication of their contributions. It is a pleasure to acknowledge the assistance of Dean Hugh Taylor, the editor of *American Scientist*, and the staff of the Yale University Press

Wallace R. Brode

Washington, D.C.
December 1966

ACADEMIA AND INDUSTRY—THEIR MUTUAL INFLUENCE*

By Hugh Taylor
Editor-in-Chief,
American Scientist

"Let men look ahead to a time when scientific efforts will [illegible]

Bacon, scientists will begin to show their strength."

Catherine Drinker Bowen†
Francis Bacon—The Temper of a Man

The establishment of the Scientific Research Society of America some sixteen years ago and its generous encouragement and endowment by Mr William Procter is significant of the changes that have been occurring during the present century in the relations between science originating in academic institutions and the researches, in ever-increasing volume, from industrial centers Two decades ago, the Society of the Sigma Xi, while rapidly expanding its activities in chapters and clubs, was forced to recognize that there were "Companions in Zealous Research," of merit and distinction, who had never had the opportunity of election to Sigma Xi but who, from the nature of their present occupations and the character of their scientific discoveries, were well qualified

*The 1964 Procter Prize award address

to share in the encouragement of scientific research, the principal objective of the Sigma Xi Society. In the colleges and universities of the country, the parent Society had attained a prestige, paralleling that of the older Phi Beta Kappa Society, but limited to professors and students in the scientific fields. In order not to alter the academic background of the Sigma Xi organization it was decided, rightly I believe, to establish a sister society with similar objectives but operating in the industrial and government research centers of the land. At the last annual convention, in December 1963, it was reported that RESA now has eighty-five branches installed or authorized, a membership of about 11,000, and an endowment with a market value of some $200,000. For a young lady in her teens, RESA (may we call her a daughter society of Sigma Xi?), with such a splendid dowry, gives a very attractive and impressive promise of adulthood.

Meanwhile, we wish to gain some perspectives on and insight into the mutual relations between academia and industry, private and governmental, as they have developed since the industrial revolution initiated the rate of change in scientific development. That revolution has become, in these latter decades, a scientific revolution, of chain characteristics, resulting in an explosive expansion of technology, an intrusion into the lives of everyone, and a scope extending to the outermost regions of space and time. Alexander sighed for new worlds to conquer. Science and technology have not only explored the moon but have received and understood messages about events that occurred before man was.

Catalysis

It is convenient to use as a launching pad for an excursion into the roles of academia and industry the subject of cataly-

sis and catalytic reactions. It was the researches in the first decades of the nineteenth century, largely in university and institutional centers, which uncovered the activity of finely divided platinum in a variety of oxidation processes Edmund Davy, Dobereiner, Sir Humphry Davy, Erman, Turner, Henry, and Faraday are all associated with this activity. Dulong and Thenard in France showed that gold, silver, and even glass shared the same property if the temperature of these agencies was sufficiently raised Faraday's researches provided insights into the mechanism of catalytic change and into the causes of "poisoning," contributions that remained unmatched well into the twentieth century. Berzelius, in 1835, coordinated a number of isolated observations and outlined his ideas about a "catalytic force" which Liebig regarded as superfluous, substituting a hypothesis of "molecular vibrations," unassailable because it could not be submitted to experimental test

A Bristol vinegar merchant, Phillips, in 1831, was the first to attempt to use platinum industrially for the oxidation of sulfur dioxide His material rapidly poisoned and the process was abandoned Squire and Messel, in 1875, made the process a technical success, using platinum catalysts to convert sulfur dioxide from pure sulfuric acid into "oleum," sulfuric acid containing dissolved sulfur trioxide. The industrial production of oleum by the contact sulfuric acid process was the achievement of German industrialists at the end of the nineteenth century, they solved the problem of removing poisons from the sulfur dioxide obtained from pyrites, and thus paved the way for the massive development of the dyestuff industry, with synthetic indigo as the initial industrial target

The honors in the field of hydrogenation and dehydrogenation go to Sabatier and his co-workers in the university laboratories of Toulouse Nitrogen fixation to yield ammonia stemmed from the classical researches of Haber in

Karlsruhe and Berlin. The Badische Anilin u. Soda Fabrik had transformed the thermodynamic and equilibrium studies of Haber on a variety of catalysts to a technical process, ready in May 1914, an achievement that freed Germany from dependence on Chilean saltpeter and led ultimately to the penetration of synthetic ammonia into the explosive and fertilizer industries of the whole world. In 1925, the same German technical organization adapted the discoveries of Sabatier to the production of synthetic methanol on oxide catalysts, the hydrogenation–dehydrogenation characteristics of which Sabatier had amply demonstrated. His studies of metal hydrogenation catalysts, such as nickel, led directly to the industrial development of fat-hardening.

Major basic scientific contributions to the study of reactions at surfaces came from the work of Langmuir in the laboratories of the General Electric Company at Schenectady, work which provided research activities far and wide in university centers such as Cambridge, Bristol, and Princeton in the 1920s. During the same period, the Fixed Nitrogen Research Laboratory in the United States Department of Agriculture entered the field of basic research on the synthesis of ammonia. Doubly-promoted catalysts, slow sorption of nitrogen as a rate-determining step, and the definition of surface area by the Brunauer-Emmett-Teller method are among the outstanding contributions from government laboratories to this field.

Reviewing the whole story, one has to conclude that academia provided the initial impetus, the continued injection of new ideas and concepts. Industry contributed magnificently to technical development, and government laboratories supplied much that neither industry nor the academics had found.

Were we to extend this survey to the developments of the last thirty years, especially in the petrochemical field, in the production of polymeric materials, rubbers, and plastics, the same pattern of interactions could be observed.

Radio Astronomy

Just over thirty years ago one of the newest sciences was born, as a direct result of communications research in the laboratories of the Bell Telephone Company. When Karl Jansky set up in Holmdel, N J., in the fall of 1930, a 14.6-meter rotatable antenna and its associated receiving and recording equipment, he listened to static from local thunderstorms, static from thunderstorms some distance away, and static from a third group "composed of a very steady hiss static the orgin of which is not yet known." In April 1933, at a meeting of the International Scientific Radio Union, Karl Jansky presented a paper on "Electric Disturbances Apparently of Extra-terrestrial Origin " We recognize the scientific modesty behind the use of the word "apparently" and we know that Jansky was receiving his messages from a large elliptical area which is aligned with the general direction of the Milky Way [1]

We have been witnesses to the birth and development of a new science that originated in an industrial laboratory Within one generation the new science had penetrated the universities, and discovery followed discovery with bewildering speed. The reach of our senses has been enormously magnified Radio stars were found beyond the limits of the astronomical instruments Exploding stars, galaxies in collision, spiral arms in our local galaxy, and signals from neighboring galaxies and from the sun, Jupiter, and Venus, powerful radio-emitters, have been found Around the world, parabolic reflectors are focusing signals not only from the outermost regions of space but also from the efforts of man, in satellites around the earth and expeditions around and onto the moon and other planets Jodrell Bank monitored *Ranger VII* until the moon passed below the horizon just before impact Higher resolutions and much greater information can be obtained by using a radio-cross The Astronomy Center of the University of Sydney is now building a large

cross-type radio-telescope at the Malonglo Radio Observatory some twenty miles from Canberra. The arms of the cross are approximately one mile in length and forty feet wide. At present the east–west arm has been completed. The north–south arm poses more difficult technical problems and, it is hoped, will be completed soon. Radiation has been detected over a small waveband at 21 cm, an emission line produced by the neutral hydrogen atom concentrated within the spiral arms of our galaxy and in other galaxies besides our own[2]

> Studies in radio astronomy range in wave length from perhaps 1 cm to about 30,000 cm . . . this is roughly 12 octaves. Previously the wave length range available to the astronomer extended from about 3,000 to perhaps 30,000 Å, or roughly 4 octaves. Between the two lies a vast region of about 12 octaves.[1]

Southworth summarizes, in conclusion, the achievements of the radio astronomer in these words:

> Following Reber's very creditable start at mapping the radio heavens, the astronomer is now filling in a substantial amount of detail. With his new tools he is having quite as much of a field day as did Galileo more than three centuries ago when, with his newly invented telescope, he discovered in rather rapid succession, the moons of Jupiter, the rings of Saturn, and the dark spots of the sun's disk as well as certain mountainous features of the moon.

Nuclear Science

The development of nuclear science has been confined almost exclusively to academia from its beginning in the

last decade of the nineteenth century to the time when Fermi and his group first demonstrated the existence of a nuclear chain reaction.

The discovery of X rays by Roentgen in 1895 was followed rapidly by Becquerel's discovery of radioactivity and by the demonstrations of the existence of the electron in the last five years of the nineteenth century. In these same years, the Curies were revealing the presence of radium and of polonium in separations from pitchblende. Rutherford showed the complex nature of the radiations, the first two components (the α- and β-rays) having different penetrating powers Pierre Curie found a radiation which was not deflected in magnetic fields and was designated γ-radiation. By 1908, Rutherford and Royds had proved that the α-particles were helium nuclei, which became the probes for the study of atomic structure Alpha-particle scattering led Rutherford in 1911 to suggest the modern picture of the nuclear atom Soddy, Russell, and Fajans were studying the radioactive elements and their transformations from which resulted, in 1913, the concept of isotopes In the same year, J J Thomson demonstrated the existence of such isotopes in nonradioactive elements, a domain which Aston richly explored from 1919 onward Also in 1919 Rutherford, bombarding nitrogen with α-particles, changed nitrogen nuclei into oxygen nuclei.

Researches by Bothe and Becker in Germany, and by the Joliots in Paris, were the preliminaries to the proposal by Chadwick in 1932 of a new radiation consisting of uncharged particles with the approximate mass of the proton The existence of neutrons was experimentally established and became the newest tool in the production of new isotopic and radioactive species because of its ability to penetrate the charged nuclei of atoms The development of particle accelerators by Van de Graaff, Cockcroft, and Walton, and by Lawrence placed other tools at the disposal of the scientist

for the exploration of the nucleus and its transformations. Einstein's statement of the generalized equivalence of mass and energy provided the quantitative bases for the energies required for or produced by nuclear transformations. It was, however, the interaction of neutrons and uranium nuclei, initiated by Fermi in 1934, which culminated in the observations of Hahn and Strassman in 1939 that the products of interaction were nuclei approximately one half the mass of the original nucleus, one of the products being barium. Meitner and Frisch, in Copenhagen, realized the significance of the observations and were aware of the immense energy that could be released in the fission. This energy release was confirmed immediately in many laboratories throughout the scientific world. It led to the demonstration of the nuclear chain reaction by Fermi and his group on December 2, 1942, "in secret, in a war laboratory, heavily financed by the United States" as H. D. Smyth observed.[3] Nuclear science passed out of the hands of the academicians; industry and governments took over. To continue, in the words of Smyth:

> In many respects the discovery of uranium fission marks the end of an era in scientific research. It was truly international, it was made by small groups working on a small scale, for the most part in university laboratories, and it was made in the atmosphere of freedom and frankness that had meant so much to science. It remains to be seen how fully we can return to such conditions.

Waves and particles. In 1913, Niels Bohr published his work on the hydrogen atom, with the electron in stationary states revolving around the positive nucleus, the proton, absorption and radiation involving quanta, $h\nu = E_2 - E_1$, where ν is the frequency, h is Planck's constant, and E_1 and E_2 are the energies in the two states. The experiments of Franck and Hertz verified the prediction of discrete energy levels in

processes involving collisions of electrons with atoms, but attempts to extend Bohr's ideas to more complex nuclei than hydrogen presented increasing difficulties met empirically with more and more quantum numbers until, finally, they gave way to the new quantum mechanics of Schroedinger and Heisenberg in 1925-26. By 1930, the application of the new principles to the electronic structure of atoms and their chemical behavior was well advanced.

It was Louis de Broglie, in 1924, who paved the way for the new wave and quantum mechanics He reversed the customary mode of thought concerning the atomistic discrete structure hitherto characteristic of matter. As with radiation, with wave properties and quanta, de Broglie proposed that material particles could also possess the wave nature of light This development is pertinent to our present theme, since confirmation of the wave nature of electrons was obtained in 1927 by Davisson and Germer in an industrial research laboratory from a study of the reflection of electrons from a single crystal of nickel G P. Thomson, in the same year, provided the additional evidence required, by demonstrating that a stream of accelerated electrons, passing through thin films of metal, produced characteristic diffraction patterns, the radius of the diffraction ring being proportional to the wavelength λ of the electrons diffracted, a wavelength given by the de Broglie relationship,

$$\lambda = h/mv$$

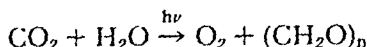
with a particle of mass m and velocity v, h being again Planck's constant Thomson has recently generously pointed out that the earlier work of Davisson in industry led him to his experiments in academia

Tools and Techniques in Scientific Research

It would be impossible in the course of one address to survey the respective contributions of university, govern-

ment, and industrial research in all branches of scientific effort. As an alternative, one can summarize the tools and techniques that are employed in modern science and show how great is the mutual influence of academic and industrial science in these areas.

A typical story in such development is to be found in the area of photosynthesis. Fifty years ago, an examination question on this topic could be adequately answered by a basic overall equation:



$$CO_2 + H_2O \xrightarrow{h\nu} O_2 + (CH_2O)_n$$

The examinee would receive high marks if he put $h\nu$ over the arrow connecting reactants and products. How research in the intervening years has progressed to the point that Calvin could receive the Nobel Prize in Chemistry in 1961 for his researches in this field can be well illustrated by an excerpt from a recent book by Calvin and Bassham on this subject. Figure 3 in their book reproduces the results of a radioautograph of a two-dimensional paper chromatogram of an alcoholic extract of the alga, *Chlorella pyrenoidosa*, after 10 seconds of photosynthesis with radioactive carbon dioxide $C^{14}O_2$. Identified in the chromatogram are alanine, malic acid, aspartic acid, phosphoenolpyruvic acid, triase phosphates, 3-phosphoglyceric acid, sugar phosphates, and sugar diphosphates. The experiment involved the development of chromatography, the use of radioactive isotopes as a tracer tool, and autoradiography to determine the extent and mode of entry of radioactive carbon into the photosynthetic cycle. As long ago emphasized by Newton, the new scientist surveys farther horizons from the shoulders of those who have preceded him.

A random selection of advances over the years can be cited in illustration of tools and techniques stemming from major discoveries in science for which industry provided superb equipment for use in future research. We can think of

Svedberg's centrifuges and their development for weight determinations of large molecules; of Irving Langmuir's utilization of the mercury vapor pump to produce high vacua; of Urey's discovery of deuterium and his no less important formulation of isotope separation involving zero-point energy differences, which gave us heavy nitrogen N_2^{15}; of Tiselius' electrophoresis apparatus; of Martin and Synge demonstrating the principles of chromatography with their instrumental developments, with columns, paper, and thin-film chromatographs; of Bridgman's high-pressure techniques that, combined with temperatures reaching to 9,000°F for brief intervals or to 5,000°F for hours, resulted in the production of industrial man-made diamonds by the General Electric Company. Progress in solid-state physics and semiconductors came largely from the industrial laboratories, and the findings of Shockley, Brattain, and Bardeen in the laboratories of the Bell Telephone Company gave the world the transistor and its manifold applications in all areas of science and life. Libby's technique of carbon dating gave a new tool to archaeological and allied geochronometrical studies. Cockcroft, Walton, and E. O. Lawrence did the pioneering work that has led to the development of the multibillion-electron volt machines now in use in the exploration of the nucleus. Powell's observation of high-energy particle tracks in emulsions and Glaser's bubble-counters are essential complements to such work in the nuclear field. The Mossbauer effect now takes its place among those yielding instruments for the elucidation of chemical structures, notably in the newly investigated noble gas compounds. Nor must one overlook the more mundane spectroscopic instruments which in the infrared, visible, and ultraviolet provide invaluable facilities for the elucidation of complex organic structures. Instrument makers have contributed overwhelmingly to the rapid advances of modern science both in the universities and industry.

I forbear, for reasons of colossal ignorance, to say anything about the computer revolution now twenty years old. In the Bell Laboratories Computing Centers, for example, the demands for computing services have grown exponentially since 1950, doubling in about thirteen to sixteen months. While leadership in computer science and technology has now reverted to the industrial laboratory, there are memorable names from academia associated with the development. One thinks initially of the English logician R. Turing in 1927, of Vannevar Bush and his differential analyzer, of Howard Aiken at Harvard and his computing machines, of John von Neumann's interest in ENIAC in Philadelphia and JONIAC at the Institute for Advanced Study and his posthumous book *The Computer and the Brain*. Industrial research "has now accomplished some of von Neumann's most daring aspirations, such as the design of computers by computers, the virtually complete solution of the mechanization of algebra, etc."[4]

Still more recent is the development of masers and lasers, now a bare ten years from conception, with academic origins through the studies of Townes, of Weber at the University of Maryland, and in Russia.[4] The patent of Schawlow and Townes, granted to them at the Bell Laboratories and embodying their concepts as published in the *Physical Review* in December 1958 opened up the whole domain of optical masers and lasers, the development and application of which in science and in industry is now in full cry.

A Personal Glance Backward

In 1917, Captain E. K. Rideal had been withdrawn from behind the Somme trenches where daily he had dosed water supplies with chlorine for the troops in his area. In the Munitions Inventions Department in London we were studying the catalytic conversion of water gas and steam

to give a cheaper hydrogen supply. We used a flue-gas analyzer for carbon dioxide which metered the gas supply, absorbed the carbon dioxide in alkali, and then by re-metering gave the CO_2 content as a difference. At the request of the Royal Air Force, we adapted this machine to the determination of oxygen in hydrogen by catalytically converting to water and measuring the volume change. Still later, we devised a continuous recorder for carbon monoxide in hydrogen by preferential conversion of the monoxide to dioxide, absorption in lime water, and determination of the change in electrical conductivity Subsequently, in 1922, with Guy B Taylor at the DuPont Company, this method of gas analysis was generalized with a machine which dosed gas samples and absorbing liquid in suitable amounts. We began our report of this work with the statement that "Chemical industry has been singularly backward in encouraging the development of automatic devices for control of operations"

In 1921, Dr R M Burns initiated the Princeton program on adsorption of gases by catalytic materials To obtain his vacuum he used a Toepler vacuum pump, with a device of Professor G A Hulett substituting a water pump for the laborious task of raising and lowering kilograms of mercury Langmuir mercury vapor vacuum pumps do not appear in the record of Princeton adsorption studies before 1927, when Kistiakowsky used them in his work on adsorption by methanol catalysts

Harvey Neville, at a bench adjacent to Burns, was studying the interaction of steam and carbon as catalyzed by alkali carbonates He also was indebted to Professor Hulett for his method of dosage with steam This was generated by a heated wire immersed in water heated externally with steam, the electrical input to the wire determining the steam produced I am afraid that gas analyses were made by hand-operated Orsat apparatus

Early efforts to discover the nature of the adsorbed species on catalysts were made by Gauger, using electron collisions with hydrogen adsorbed on nickel, and by Wolfenden and Kistiakowsky measuring ionization potentials, the former with adsorbed hydrogen on nickel, the latter with nitrogen adsorbed on iron. Kistiakowsky attributed an 11-volt ionization potential to adsorbed nitrogen and a 13-volt potential to adsorbed hydrogen. He argued that the results indicated adsorbed atoms rather than molecules.

In a further effort to ascertain the properties of adsorbed hydrogen, studies were conducted on atomic hydrogen produced at room temperatures by photosensitization with excited mercury. Here it may be emphasized that it was necessary to build the sources of resonance radiation, one interesting form being built from two quartz-to-glass seals, the quartz ends fused together and the glass ends becoming G-702 P glass cathode and anode compartments respectively.

When the *para*-hydrogen conversion was invoked to indicate activated adsorption of hydrogen on various catalysts, it was necessary for Sherman to build his own thermal conductivity cell. Still more so, when deuterium was used as an isotopic tracer in exchange reactions, it was necessary to use ultraviolet spectroscopy as an analytical tool for deuteroammonias and deuterobenzenes and to adapt infrared spectra for analytical purposes for the measurement of deuteromethanes and alkanes. Soon a mass spectrometer became indispensable, and one was built from Nier's blueprints. There were no infrared or mass spectrograph machines to purchase at that time. It was well into the post-World War II period, when the most active mass spectrographic work in Princeton was over, that the first Consolidated mass spectrograph was purchased for Frick Chemical Laboratory.

It is interesting to note how many and varied techniques have been employed since those early days for the exploration of adsorption by catalytic materials. One thinks of

Selwood's various essays in the field of magnetochemistry, and of proton relaxation and catalyst accessibility; of Turkevich's studies in electron microscopy and small-angle X-ray scattering; of Eischen's studies of infrared spectra of adsorbed species with, now, an extended bibliography from centers throughout the scientific world; of Beeck's techniques of thin films of catalyst metals which, since his death, have yielded a rich harvest to Kemball, Tompkins, and others; of radioactive tracers, principally of tritium and carbon-14, which have helped to delineate the catalyst surface; of gas chromatography which permitted Emmett to perform, in minutes, reactant gas analyses which in the 1920s would have occupied days, or permitted Beebe and others to explore by alternative techniques the basic problem of adsorption equilibrium, nuclear magnetic resonance and electron spin and paramagnetic resonance have been brought into play; semiconductivity and catalysis by compounds and *d*-band holes in metals and alloys have led to electronic interpretation of catalysis All this magnificent development in just one area of scientific research has been made possible largely by the contributions of industrially developed instrumentation in the postwar period and has enormously accelerated the pace of progress This is probably the most significant impact of industry on academia in scientific research

I end on a personal note In 1925, by a deductive leap in the dark, I suggested that catalyst surfaces could not be homogeneous, that from the heterogeneity might stem some of the most characteristic properties of catalytic materials Active centers became the focus of argument "about it and about" There were no tools then to force a decision But in the intervening years, as Ehrlich of the General Electric Research Laboratory[5] pointed out a year ago:

> The outstanding feature of the past decade has been the development and perfection of a variety of experi-

> mental techniques that allow a new and deeper insight into the structure of surfaces, and into the elementary atomic events occurring on them ... What is important is that experiments properly executed and interpreted, can now remove from the realm of speculation most of the elementary facts dealing with interplay of gases and an initially clean surface. This tremendous advance has, in large measure, been made possible by the reduction of ultrahigh vacuum techniques to a matter of complete routine.

Muller's field ion microscope resolves the details of the atomic arrangement of the surface, and Ehrlich has shown that direct observation of adatoms is indeed possible..

W. O. Baker[4] has informed me, on this same topic, that low-energy electron diffraction studies conducted by Germer and MacRae at the Bell Telephone Laboratories have explicitly confirmed the prediction of activated surfaces and that, in the same laboratory, Lander has shown differences in the spacings and thus in the valence bonding of silicon atoms on especially prepared cleaved surfaces of the element, employing also low-energy electron diffraction. Industry thus makes academic dreams come true.

The Future

What does this peering into the past tell us about the future? Long ago I said that the mantle of the prophet lies uneasily on the scientist. I. I. Rabi in an article in the *New Scientist*[6] expressed it more pungently. "It is a characteristic of scientists in general that they have no flair for predicting the future. That is better done," he wrote "by the H. G. Wellses and Aldous Huxleys."

We can, however, indicate trends in academic, industrial, and governmental relations which will surely emerge in

the coming decades We can be certain that the present exponential growth in science that has characterized the post-World War II years will continue until economic and other factors produce a developmental plateau. The relations between academia and industry will become more intimate and interrelated We can expect the lag between discovery and application, which formerly was some twenty years but which has shortened so dramatically in recent times, may well be further abbreviated. The explosive increase in the number of scientists makes certain that the tempo of application will remain high.

It has been in the universities that the "seed corn" of technological progress has often been raised during the past century This makes *it imperative that, in the universities,* there exist a climate favorable to the production of basic science from which technological applications inexorably result That climate requires leisure time to think, on the part of first-rank scientists, unhampered by undue calls on their services as consultants to industrial and governmental agencies, to the maximum degree possible Such scientists must be free to roam where their scientific spirit moves them, again unhampered by calls for the solution of particular scientific and industrial problems. To that end, we might well consider a major extension of governmental, foundational, and industrial financial support to key scientists in the universities across the land to pursue *their own* scientific objectives, without regard to particularized contractual obligations and requiring only the reports on their endeavors that emerge in their scientific publications By way of example, one can note the annual reports on researches, published by research professors, research fellows, and research students, that appear annually in the *Year Books of the Royal Society of London* Endowments approaching $5 5 millions permit expenditures by the Royal Society in excess of $200,000 per year for such purposes In addi-

tion, the Society administers parliamentary grants for research professorships and scientific investigations, exceeding $350,000 in the year 1963-64. The researches so supported are characterized by their basic scientific interest and their freedom from anything other than personal choice. These choices range from molecular biology to X-ray spectroscopy and radio astronomy, from Mendelian populations to high-energy physics and satellite launching, from cavitation to geochronometry.

Industries on the eastern and western seaboards of this continent have already demonstrated the immense significance of the pursuit of basic science in their own research laboratories, and its fruitfulness. Until a similar condition obtains in the areas between these rich scientific technological concentrations, the potentialities of the country are not being fully realized. To attain progress in the central areas, consideration might well be given to the establishment of research institutes now in successful operation in the south, the southwest, and in Stanford, to name just a few, where a central research institute, competently staffed by skilled personnel, can count on the support and cooperation of interested industries. For the scientifically underdeveloped areas of the country some consideration might be given to plans for and probable future of such research institutes. A current example from Canada has been recently discussed by Dr. A. D. Misener, Director of the Ontario Research Foundation,[7] a plan to strengthen Canada's technology, to lessen Canada's reliance on foreign technology, and to halt the "brain drain" which has resulted from the lack of research opportunities. A versatile research community is planned for a 339-acre site near Toronto. The Foundation will occupy 100 acres of the plot. The remaining acreage is is to be allocated to industrial companies that wish to participate. The International Nickel Company of Canada, Consolidated Mining and Smelting Company, Dunlop Inter-

national Research, and British American Oil Company have agreed to join and purchase land The Province of Ontario is giving both financial assistance to and encouragement of the project. Sheridan Park will constitute a challenging enterprise which might well be emulated in areas south of the Canadian border.

The gross national product in the United States has been rising by 25 to 30 billion dollars annually in recent years, and will soon surpass, if it has not already passed, the 600 billion dollar mark What fraction of this huge productive effort should be returned for the further support of science and technology, with which academia and industry may ensure the future progress of the country, is a question that should engage the serious consideration of all who are qualified to contribute to the answer It will be an important answer for the welfare of the country

In accepting the William Procter Prize for 1964 I wish to acknowledge all the help, kindness, and consideration from students, colleagues, and friends that alone have made this award possible

REFERENCES

1. C M`Jansky, Jr , *Am Scientist, 45*, 5 (1957), G C Southworth, *Sci Monthly* (Feb 1956)
2. B Y Mills, *New Scientist, 23*, 570 (1964)
3. H D Smyth, *Am Scientist, 35*, 485 (1947)
4. W O. Baker, personal Communication
5. G Ehrlich, *Advan Catalysis, 14*, 225 (1963)
6. I. I Rabi, *New Scientist, 21*, 137 (1964)
7. A D Misener, *New Scientist, 23*, 636 (1964)

2. LIQUID HELIUM

By Lyle B. Borst
State University of
New York at Buffalo

Liquid helium has amazed physical scientists for more than three decades. As a substance, it is so remarkable that a new *category called quantum liquids was established, of which it* is the only member. By examining liquid helium, we see a new world, different from the one we know, where quantum phenomena are major effects and are not confined to the submicroscopic world of atomic physics. We perhaps have our first peek into "looking-glass house," where things are topsy-turvy. Superfluid helium may be the first material in which the atom plays no part. Superfluid helium may be an element in the Aristotelian sense which is continuous and has no structure.

Helium, as a gas, shows no surprising characteristics inconsistent with its low molecular weight.[1-3] It will permeate substantial objects, the more rapidly the higher the temperature, but this is true in varying degree for hydrogen, neon, and other light gases. Solid helium, likewise, has relatively few characteristics that are unexpected. But liquid helium is a different matter entirely.

The size of gas molecules has long been measured by determining the unavailable space during compression. When the free space becomes sufficiently small, the gas condenses to a liquid (if the critical temperature is not exceeded). The size of a molecule of liquid is what one expects from the gas

measurements. In liquid helium, however, the density of the liquid is a quarter of that expected, so the molecule (or atom, since helium is monatomic) has four times the expected volume. The wide spacing of the atoms cannot be explained by classical physics and is considered primary evidence of quantum effects.

In 1923 De Broglie asked: If light waves have particle properties, can material particles not have wave properties? De Broglie waves have been the foundation for modern wave mechanics. In the case of helium, De Broglie waves, extending beyond the atoms, overlap and prevent the atoms from approaching each other closely. This wave-overlap prevents liquid helium from freezing. In this respect helium is unique, since it is the only liquid that does not freeze. In order to form crystals, helium atoms must be forced together until the intermolecular Van der Waals forces predominate and form the ordered lattice. Even at temperatures of less than 1°K, solidification occurs only at 30 atmospheres pressure.

Liquid helium at its normal boiling point of 4.2°K is interesting but not remarkable. It is transparent, and colorless, is a poor conductor of heat, and is an electrical insulator. It has a low viscosity and a low heat of vaporization. It is a vexing substance to handle but not remarkable. As the temperature is lowered, it contracts, as do most substances, until the temperature of 2.18°K is reached. At this temperature, the liquid suddenly expands. The rate of expansion diminishes as the temperature is further lowered until, near 1°K, the rate is zero and the normal contraction returns. The transition at 2.18°K is sudden and comprehensive, and marks the boundary between classical and quantum fluids.

The very nature of liquid helium changes at the lambda point (2.18°K). Although the liquid is a poor conductor of heat above this temperature, below it, it becomes the best conductor we know—500 times better than copper or silver! Generally, good conductors of heat are also good conductors

of electricity, but helium has no free electrons and remains an electrical insulator. Thermal conductivity in helium must be of a completely new variety; for its description, a new phenomenon is identified: second sound Above the lambda point, liquid helium is constantly boiling, since even in the best apparatus some heat penetrates. Below the lambda point, all bubbling stops for the heat is conducted to the free surface where evaporation takes place This transport of heat is best described by the equations of acoustics; and although no one will ever hear "second sound," the term is appropriate. Above the lambda point, liquid helium can be stored indefinitely in a Pyrex vacuum flask Below the lambda point, however, the vacuum soon deteriorates because helium creeps through the glass. Helium-II, the low temperature form, is a superfluid It penetrates cracks and holes that no other material can find The unwary apparatus designer who leaves a void in his apparatus accessible by superleak, will have an explosion when the temperature rises, for pressures of thousands of atmostpheres can easily develop.

Attempts to measure the viscosity of the superfluid lead to contradictory values Andronikashvili cleverly demonstrated these anomalies by studying the rotation of a series of closely spaced disks immersed in the liquid His apparatus consisted of a vertical shaft upon which the disks were mounted, supported by a torsion fiber, so that the system could oscillate much as the balance wheel of a watch. He noted the period of oscillation in air and, knowing the characteristics of his fiber, determined the moment of inertia of his system. If the same measurement were made in a viscous liquid (like salad oil) the moment of inertia would be that of the shaft, the disks, and the oil between disks In helium above the lambda point, the liquid was carried by the disks and showed the expected density At a temperature near 1°K, however, no helium was carried. Apparently the liquid showed zero viscosity, for it did not reflect the motion of the nearby disks As the

measurements. In liquid helium, however, the density of the liquid is a quarter of that expected, so the molecule (or atom, since helium is monatomic) has four times the expected volume. The wide spacing of the atoms cannot be explained by classical physics and is considered primary evidence of quantum effects.

In 1923 De Broglie asked: If light waves have particle properties, can material particles not have wave properties? De Broglie waves have been the foundation for modern wave mechanics. In the case of helium, De Broglie waves, extending beyond the atoms, overlap and prevent the atoms from approaching each other closely. This wave-overlap prevents liquid helium from freezing. In this respect helium is unique, since it is the only liquid that does not freeze. In order to form crystals, helium atoms must be forced together until the intermolecular Van der Waals forces predominate and form the ordered lattice. Even at temperatures of less than 1°K, solidification occurs only at 30 atmospheres pressure.

Liquid helium at its normal boiling point of 4.2°K is interesting but not remarkable. It is transparent, and colorless, is a poor conductor of heat, and is an electrical insulator. It has a low viscosity and a low heat of vaporization. It is a vexing substance to handle but not remarkable. As the temperature is lowered, it contracts, as do most substances, until the temperature of 2.18°K is reached. At this temperature, the liquid suddenly expands. The rate of expansion diminishes as the temperature is further lowered until, near 1°K, the rate is zero and the normal contraction returns. The transition at 2.18°K is sudden and comprehensive, and marks the boundary between classical and quantum fluids.

The very nature of liquid helium changes at the lambda point (2.18°K). Although the liquid is a poor conductor of heat above this temperature, below it, it becomes the best conductor we know—500 times better than copper or silver! Generally, good conductors of heat are also good conductors

of electricity, but helium has no free electrons and remains an electrical insulator. Thermal conductivity in helium must be of a completely new variety; for its description, a new phenomenon is identified: second sound. Above the lambda point, liquid helium is constantly boiling, since even in the best apparatus some heat penetrates Below the lambda point, all bubbling stops for the heat is conducted to the free surface where evaporation takes place. This transport of heat is best described by the equations of acoustics; and although no one will ever hear "second sound," the term is appropriate. Above the lambda point, liquid helium can be stored indefinitely in a Pyrex vacuum flask. Below the lambda point, however, the vacuum soon deteriorates because helium creeps through the glass Helium-II, the low temperature form, is a superfluid. It penetrates cracks and holes that no other material can find The unwary apparatus designer who leaves a void in his apparatus accessible by superleak, will have an explosion when the temperature rises, for pressures of thousands of atmostpheres can easily develop

Attempts to measure the viscosity of the superfluid lead to contradictory values Andronikashvili cleverly demonstrated these anomalies by studying the rotation of a series of closely spaced disks immersed in the liquid His apparatus consisted of a vertical shaft upon which the disks were mounted, supported by a torsion fiber, so that the system could oscillate much as the balance wheel of a watch He noted the period of oscillation in air and, knowing the characteristics of his fiber, determined the moment of inertia of his system. If the same measurement were made in a viscous liquid (like salad oil) the moment of inertia would be that of the shaft, the disks, and the oil between disks In helium above the lambda point, the liquid was carried by the disks and showed the expected density. At a temperature near 1°K, however, no helium was carried. Apparently the liquid showed zero viscosity, for it did not reflect the motion of the nearby disks. As the

temperature was raised, various fractional densities were observed until the full density was reached at the lambda point. The term "normal component" was applied to the fraction of the liquid that rotated, whereas the remainder was termed the superfluid component.

If two branches of a U tube be connected by a capillary and if one branch be filled higher than the other, the liquid levels will oscillate over minutes or hours until they become equal. No such phenomenon is known except in liquid helium below the lambda point. The explanation given is that the normal component cannot pass the capillary and that the superfluid passes back and forth, changing the composition and the temperature. If, then, one branch is heated, the composition of the fluid will change and the superfluid will move toward the high-temperature side, producing a pump. This appears to be a violation of natural laws akin to perpetual motion. When, however, the applied heat is included in the thermodynamics, the system becomes a heat engine, without working parts, in which heat is converted directly to mechanical work. Now if the hot branch has a restriction, hydraulic pressure is produced and liquid is projected to form a fountain when the liquid reaches this point. This fountain continues to play as long as there is helium in the system and the two temperatures are below the lambda point and are unequal.

When an open-topped container filled with liquid helium is suspended above a pool of liquid (always below 2.18°K) it will slowly empty itself. This emptying is not due to diffusion through the wall. This happens in any shape container made of any material. The liquid wets any solid, creeps up the wall, and overflows to drip off the bottom. Again, this is a quantum phenomenon not found in other substances. The "velocity of creep" is nearly independent of the nature of the container and its height, the rate of flow depending primarily on the wetted perimeter.

In contrast to helium-4, the rare isotope helium-3 shows

none of these remarkable superfluid characteristics. Helium-3 was first studied in order to throw light on helium-4; however, it soon showed such remarkable behavior that it is now studied in its own right. In helium-3 there is an anomaly in thermal expansion but of a very different variety Helium-3 is paramagnetic Although helium-4 and helium-3 are isotopes, they are more unlike than any other pair of isotopes. In fact, they are so different that, when a solution of equal parts is cooled below 0 9°K, the liquid separates into two phases The heavier, the helium-4-rich phase, shows superfluid properties, somewhat suppressed. The light helium-3 phase shows no superfluidity Again these differences are of a quantum nature Helium-4 has even particles in its nucleus, two protons and two neutrons, whereas helium-3 has odd particles, two protons and one neutron Odd and even numbers in quantum mechanics are given the most profound importance The statistics first proposed by Bose and Einstein apply to light and generally to systems having even characteristics Fermi and Dirac developed statistics particularly appropriate to the electron and to systems with odd characteristics Helium-4 uses the mathematics appropriate to radiation problems, and helium-3, the mathematics of electric systems

Quantum mechanics is usually considered as a highly mathematical branch of theoretical physics divorced from everyday experience Historically, this has certainly been true, still, there are quantum problems in our world which have the salient quantum features and which differ primarily in scale from atomic systems. As an illustration, let us take the problem of traffic in a city In order to simplify the problem we shall assume the blocks to be square and of identical size. A common size is eight blocks to the mile Now let us assume that there is a traffic light at every intersection All lights turn green in a north–south direction simultaneously and 30 seconds later turn green east and west. At what speed can cars

travel in any direction without being stopped by a red light? If a man driving north covers a block in just one minute, lights will turn green as he approaches, however far he drives. His velocity will be:

$$v = d/t = 1/8 \text{ mile per minute} = 7.5 \text{ mph}$$

Southbound cars can travel at this speed also and so can cars traveling east and west!

Is this the only solution to the problem? A car driving faster than 7.5 mph will inevitably be stopped by a red light. If it travels at half this speed, it will cover a block in two minutes. At one-third the speed, in three minutes, etc. So the complete solution is

$$v = \frac{7.5}{n} \text{ mph}$$

with n an integer. The quantum number n is a consequence of the uniformity of the street grid. Quantum numbers in atomic problems appear in like manner whenever periodicity occurs.

Such quantum numbers also occur in acoustics. In the case of the violin string they are called harmonics. If the string vibrates as a whole then $n=1$ and the tone will be the fundamental. Another tone an octave higher can be obtained by damping the string at its midpoint. The string vibrates in the form of an S with stationary midpoint and $n=2$. The third harmonic is an octave and a half above the fundamental. In this case the string vibrates in three sections. So, n can take any integral value. In the case of a string of beads, however, n can take values only to the number of the beads, N, for the Nth harmonic will represent adjacent beads oscillating in opposite directions.

The timpani or kettledrum is a two-dimensional vibrating membrane that shows similar quantized overtones involving two independent quantum numbers.

Vibrations of a three-dimensional solid require three

quantum numbers n_x, n_y, n_z, of exactly the same nature. The only musical instrument based upon vibrations of a massive solid is the musical anvil—a most limited instrument. In the atomic realm, however, the theory of acoustic vibrations of a crystalline solid gave rise to a scientific achievement which will stand through the centuries. In 1912 Debye studied this problem by counting the number of vibrational modes or harmonics as a function of the quantum energy of the vibration He obtained an expression for the internal energy and heat capacity which has been the cornerstone of our understanding of the solid state He determined that the heat required to raise the temperature of a given mass of solid one degree must vary as the cube of the temperature, if the temperature is sufficiently low. As the temperature rose, higher values of n occurred so that as the quantum number n approached the number of atoms N, as in the case of the beads on the string, no new vibrations were possible and the theory became that of a classical substance described by the law of Dulong and Petit

Liquid helium shows the characteristics of a Debye substance below 0 7°K and above 3°K. Between these temperatures we have the tremendous anomaly in heat capacity and thermal expansion The normal component of Andronikashvili correlates perfectly with the internal energy of the anomaly The nature of the superfluid, the nature of the anomaly, and the nature of the lambda transition are the perplexing phenomena that a satisfactory theory must explain

Current theories

Many theories have been proposed, differing in their starting point and emphasis No current theory gives a satisfactory explanation of all phenomena, so that the problem is by no means solved

The first general theory was that of Landau, proposed in 1941 and modified in 1947.[4] He considered the rotation of the whole liquid to be a quantum phenomenon and derived a relation between the energy and momentum of these quantum states or excitations. He named his excitations rotons in contrast to the acoustical excitations called phonons.[1]

In Figure 1, energy is plotted vertically and momentum horizontally. The rate of change of energy with momentum measures a velocity, so the slope of the curve represents the velocity of sound. Phonons have a velocity independent of their energy so that they are represented by the straight line through the origin. Landau's rotons show a parabolic relationship and are represented by the dotted curve.

This assumed excitation permitted Landau to fit the experimental heat capacity to 1.6°K, using three empirically determined constants. The same three constants were used to fit the normal component as well. Landau and other workers have adapted the theory to a wide variety of phenomena involving the superfluid. It has been a most constructive and useful theory and has been the basis for most experimental investigations. Generally speaking the superfluid component has a negative definition. It is the absence of rotons, i.e. the absence of the normal component. Below 0.7°K, liquid helium shows no excitations other than phonons and can be described by the Debye theory. This is also true of solid helium and of liquid helium near its normal boiling point of 4.2°K where there is no evidence of superfluidity. Landau gives no physical interpretation of the lambda point except as the temperature at which the normal component reaches 100 per cent (and the superfluid concentration becomes zero).

1. In quantum mechanics waves and particles are equivalent, so for each particle there must be a wave and for each wave there must be a particle; e.g. in electromagnetic radiation, for each light wave there is a particle called the photon; in acoustics, for each sound wave there is the particle called the phonon.

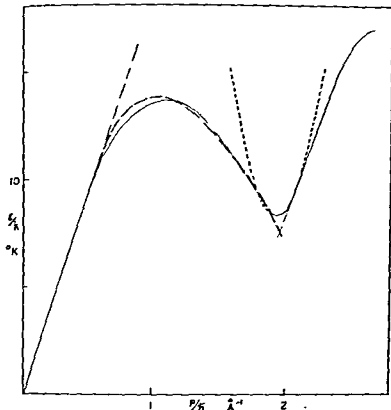


FIG. 1 Dispersion relation between the energy of an excitation and the momentum Dotted line, Landau roton, continuous line, experimental results of neutron scattering; dashed line, proposed theory

A more recent interpretation of superfluidity has been given by Feynman [5] He points out that the excitations are of two kinds, one of long wavelength involving the coordinated, coherent oscillations of large numbers of atoms (phonons), the other of higher energy and shorter wavelength involving the motions of a small number of atoms (rotons) He considers the ultimate theory derivable from the proper Schroedinger wave equation. In the absence of this equation, he tries various wave functions and chooses

that representing a vortex or smoke ring as the model having properties most like rotons. He obtains a dispersion curve (the relation of energy to momentum) similar to that found by neutron scattering. He associates the superfluid with the absence of molecular size excitations, vortices, or rotons. Again there is no physical model of the lambda transition.

Cell Theories

A new theory to justify consideration must either (*a*) give a rigorous relationship between observations and a set of well-defined postulates or (*b*) explain and relate observed phenomena in a new light, encompassing a wider range of observations and showing new relationships. The present proposal cannot yet be derived from the Schroedinger equation, although this is anticipated; nevertheless, it does interpret experimental observations quantitatively in a totally new light.[2]

De Boer[6] and Temperley[7] have independently proposed cell theories of liquid helium. Atoms are considered in pairs, free to rotate within a sphere defined by their neighbors. The two kinds of excitations are then the coordinated motions of the centers of gravity of many cells—the acoustic modes or phonons, and the rotational states of the distomic rotors within the cell. The Schroedinger equation can be expressed and solved subject to the adiabatic approximation; i.e. that rotation does not affect vibration, and acoustic vibrational modes are not coupled to the rotations within the cell. The

2. Note added in proof:

The Schroedinger equation which leads to the anomalous thermodynamic properties is that of an n dimensional harmonic oscillator. The oscillators characterize the normal modes of vibration among a set of n diatomic rotors as defined above. Two quantum numbers v and n appear in the partition function: v the vibrational quantum number and n the rotor population dependent only upon temperature (*Nature, 204*, 870, 1964).

The dispersion relation is obtained from the complete partition function including phonons (*Nature, 209*, 187, 1966).

results are the vibrational modes of Debye theory and a set of equally spaced rotational states of the diatomic rotor. This theory gives a clear differentiation between helium-3 and helium-4, since helium-3, having an odd number of particles in the nucleus, will show all rotational quantum numbers, whereas helium-4 will show only even values ($J=0, 2, 4$) The absence of rotation ($J=0$) will be the lowest state where only phonons exist and will be associated with the Debye liquid for $T \leq 0.7$°K Experimentally, the odd–even relationship accounts roughly for the fact that anomalies in helium-4 occur in the 1–2°K temperature range whereas in helium-3 they occur at 0–0.3°K. A critical examination of experimental data does not, however, disclose evidence of well-defined rotational levels The weak coupling (adiabatic approximation) between rotation and vibration would seem to be in question

The present proposal is a strong coupling theory in which vibrations and rotations are thoroughly and indistinguishably mixed Since an analytical derivation from the Schroedinger equation has not yet been obtained, an approach is taken as an extension of Debye theory In Debye's theory, the modes in the x, y, and z directions are thoroughly and completely mixed or coupled by assuming arbitrary and unrelated values of n_x, n_y, and n_z A vibrational mode is described by any three values of n The T^3 heat capacity results from this assumption Roughly speaking, the experimental heat capacity of helium below 2°K varies as the fourth power of the temperature A four-dimensional Debye theory therefore gives rough agreement with the data If now the fourth dimension is related to the rotational mode of the diatomic pair, a nonhomogeneous space is defined which acts like three dimensions at low temperature (where $J=0$) but four dimensions at higher temperatures (where $J=2$ is postulated) Complete representation of the heat capacity and of the normal component can be achieved to about 2°K.

well beyond the region of validity of the simple Landau theory (Fig. 2). The density of the liquid can be fit to 2.12°K. if an additional assumption is made that the $J=2$ state has a density 1.2 per cent greater than the Debye liquid.[3]

The range of validity of the Debye theory is limited to temperatures for which the number of acoustical modes (phonons) does not exceed the number of atoms or molecules. The present theory gives a similar set of countable states (involving rotation as well as vibration). When one equates the number of modes to the number of molecules, a temperature of 2.7°K is obtained which represents the upper bound of the theory. It is believed that 2.7°K is the theoretical value of the lambda temperature (2.18°K by measurement). The theory cannot be valid above this temperature.

The proposal requires the consideration of pairs of identical helium atoms. No permanent binding is necessary since only quantized rotation is required, the nearest neighbors defining the cell. The nearest neighbor distance is therefore assumed for the diatomic rotor. If helium-3 is added to the liquid, He^3–He^4 pairs form, thereby reducing the number of He^4–He^4 pairs. If the number of modes is now counted to the number of He^4–He^4 pairs expected, the lambda temperature is found as a function of helium-3 concentration. A comparison with experiment (Fig. 3) shows the form and the slope to be correct but, since the theory predicts 2.7°K for pure helium-4, the lambda temperatures are uniformly high. Support for the consideration of atoms in pairs results also from a study of the Andronikashvili experiment in He^3-He^4 solutions. The concentration of the superfluid component extrapolated to zero degrees will be 100 per cent for pure helium-4. In solution, these extrapo-

---

3. In the case of liquid hydrogen, the high temperature form (75% *ortho*, 25% *para*) has a density 0.5% greater than the pure *para* ($J = 0$) form. Thermal expansion will therefore not vary as T^3 but will reflect the *ortho–para* composition at equilibrium at each temperature.

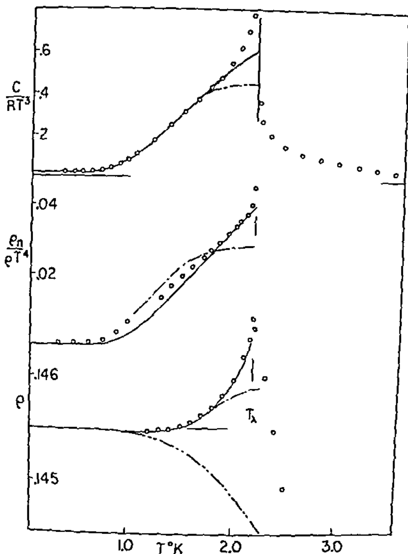


FIG 2. Temperature dependence of thermodynamic variables Circles, experimental data, continuous line, proposed theory; dot-dash, Landau theory, dash-double dot, Debye theory. Top, heat capacity, center, normal component, bottom, density.

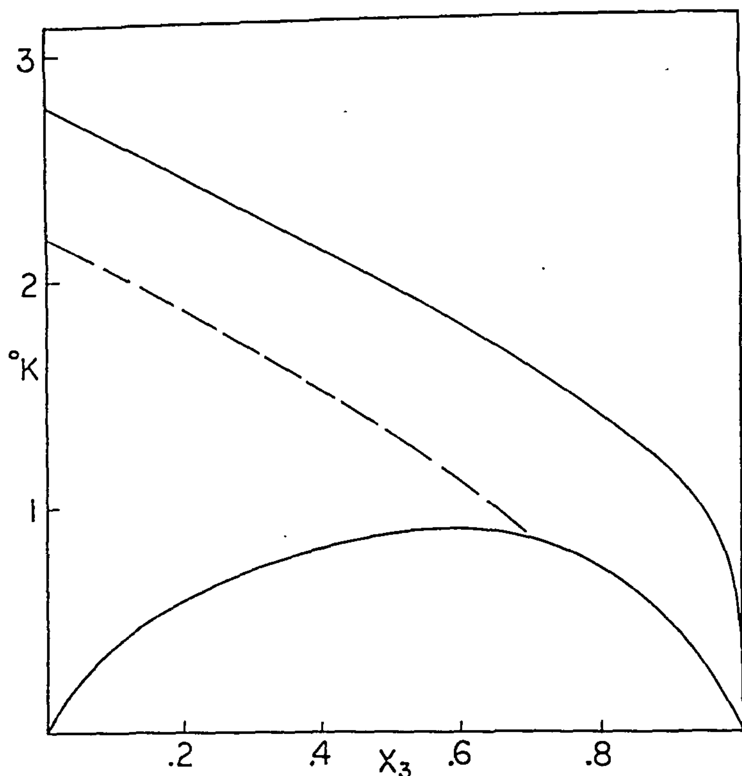


FIG. 3. Phase diagram, helium-3 helium-4 solutions. Atomic fraction helium-3 plotted horizontally. Upper line, calculated lambda temperature; dashed line, observed lambda temperature. Lower line, phase separation boundary.

lated values diminish twice as fast as would be expected for monatomic constituents, but quite in accord with the assumption of paired atoms.

There is at present no adequate explanation of the lambda point. It is considered to be a "second-order" phase transition, and is virtually unique among liquids. In solids, similar discontinuities are common and are associated with a change in the crystalline structure. Since liquids have no crystalline lattice, this explanation is not permissible. A transition of

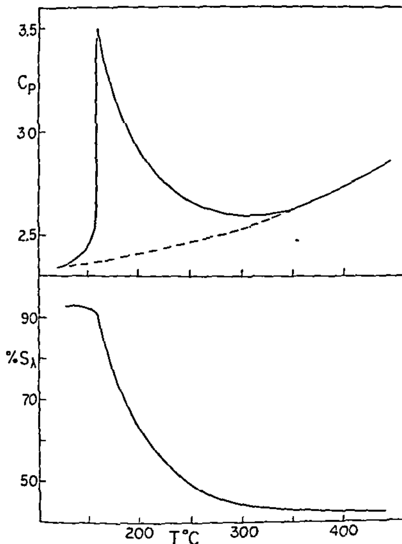


Fig. 4 Phase transition in liquid sulfur Above heat capacity, below, composition per cent low temperature form (lambda)

the same nature was observed in liquid sulfur at 160°C during the nineteenth century (Fig 4) Here, the molecular structure has been well characterized both above and below the transition temperature The transformation is one of polymeriza-

tion from an eight-membered ring (the lambda form $T < 160°C$) to a linear polymer of large molecular weight (the mu form $T > 160°C$). The heat capacity becomes large and has the same character as liquid helium although reversed in the sense of temperature. By analogy we can say that the liquid becomes increasingly ordered as the transition temperature is approached, in sulfur by increased molecular weight of the polymer, in helium by increased number of rotation–vibration coupled modes. The ordered state becomes thermodynamically unstable and transforms into the simpler form: in sulfur the eight-membered ring, in helium the monatomic liquid.

The lambda point is therefore identified as a dissociation temperature for the paired states. This occurs near the temperature at which the number of modes equals the number of molecules. Above this temperature the liquid is monatomic, again showing the character of a Debye liquid, but lacking the character of a quantum liquid. There appears to be a dissociation energy of 1.2 calories per gram – one fifth the normal heat of vaporization.

The remarkable achievements of Landau's theory stem from an assumed relation between the energy and momentum of his excited (roton) states. A major accomplishment, predicted by Feynman, was the measurement of this relation by the inelastic scattering of neutrons. An examination of Figure 1 shows why Landau's assumption was constructive, for the states near the minimum are those that have roton character. Deviations from the low-energy, low-momentum phonon states are not a part of Landau's theory but are an important aspect of the present proposal.

It is therefore important to derive the dispersion relation from the helion (strong coupled cell) theory. This is done by assuming that both the roton and helion excitations give alternative and correct expressions for the internal energy. These expressions are equated, and common factors are can-

celed to give a relation between the excitations of classical statistical mechanics and the helion. A fixed value of Debye's relative coordinate ($x=h\nu/kT$) is introduced, and the resulting expression gives an excellent fit of the observed curve up to the minimum In this formulation, the minimum corresponds to the lambda point at which the number of modes equals the number of molecules and dissociation occurs. Beyond this minimum only phonons will exist—a reasonable interpretation of the empirical data.

The theory suggests that the minimum of the dispersion curve is another expression of the dissociation occurring at the lambda temperature If this be true, then the minimum of the dispersion curve should follow the lambda temperature as a function of applied external pressure and in He^3–He^4 solution

Interpretation

Liquid helium is a Debye liquid at the lowest temperatures. As the temperature is raised, acoustical modes are excited, giving rise to the T^3 dependence of the heat capacity The De Broglie wavelength is long, compared to any characteristic dimension, and the theory can be considered that of a continuum. As the temperature approaches 1°K, the De Broglie wavelength is approaching atomic dimensions The longest constructive dimension in the system is the circumference of the orbit of rotation of an atomic pair (5.8 Å). Since the nearest neighbor distance (3.7 Å) is large compared to the size of the atom (the effective diameter of the electron cloud of the helium atom is 2.7Å) and no well-defined lattice exists in the liquid, the diatomic pair can be considered enclosed in a spherical potential defined by the nearest neighbors Quantized free rotation will then be excited after the manner of the rigid rotor The first rotational state occurs at 5 1°K, comparable to the energy of the vibrational states

(phonons) and cannot exist at equilibrium in the absence of phonons. The rotational and vibrational states are therefore assumed to be strongly coupled. As the temperature approaches T_λ, more and more pairs are excited until all atoms are participating. The number of modes then equals the number of molecules. The addition of further energy causes dissociation to the monatomic liquid, reverting to the Debye liquid with conventional heat capacity and expansion.

The superfluid appears to be a state of complete quantum-mechanical repose. At the absolute zero of temperature, no vibration or rotation occurs. At finite temperature, the superfluid is that fraction of the liquid associated with neither acoustic nor rotation–vibration modes. The De Broglie wavelength is long compared to any atomic dimension (15 Å at 1°K) and the liquid can be treated as a continuum. In this state, one can question the utility of the atomic theory, since the liquid appears to be structureless. It can penetrate openings of less than atomic dimensions. The flow of the liquid as a film may be considered as a consequence of this structureless character. An atomic liquid will have a boundary at which the density of atoms suddenly changes or becomes zero. If the liquid is structureless, a boundary is a paradox. At the edge defined by liquid, solid, and vapor, the liquid would appear to respond by creeping over the solid and the paradox of the edge of a continuum will be avoided. The Andronikashvili experiment is understandable in terms of the continuum, for viscosity is explained by a finite free path which carries atoms into the flowing fluid. If, however, no atoms are present in the continuum, no mean free path is possible and no momentum can be transferred.

This theory of excited states would seem to be general for monatomic noble gases, limited only by the freedom of the paired atoms to rotate in the cell of their neighbors. Helium-6, a radioisotope of 0.8-second half-life, would surely form a superfluid if prepared in high concentration. An estimate

indicates that the lambda point of the pure material would be expected near 1.5°K. The next heavier noble gas is neon of atomic weight 20. The De Broglie wavelength would be less than half that of helium at a given temperature, and the quantum mechanical wave interference would be expected to be considerably less The combination of smaller nearest neighbor distance and greater atomic weight gives a rotational constant one quarter that of helium, so that the Debye fluid, if it forms, would be found only below 0.2°K. The existence of a lambda point cannot be predicted with assurance, since the rotational state falls below the temperature at which Debye modes are numerous, and the requirement for strong coupling may not pertain. Moreover, the theory is that of a liquid with no long-range order, so liquid neon would have to be supercooled below its normal freezing point of 25°K If, however, this experimental difficulty can be overcome, superfluid characteristics might well be observed at temperatures of a few tenths of a degree

Conclusion

A new theory is proposed as an extension of the Debye theory in which pairs of atoms are assumed to rotate, and these rotational excitations are strongly coupled to the acoustical vibrations of the Debye theory The rotation–vibration excitations, called helions, are analogous to the rotons of Landau and the vortices of Feynman They account quantitatively for the thermodynamic anomalies in liquid helium, predict a lambda point, account for the lambda point in He^3–He^4 solutions, and give a dispersion relation of the observed form The superfluid is that fraction of the liquid corresponding to neither acoustic nor vibration–rotation excitations, or that fraction in the zero momentum or ground state As such, the superfluid is a continuum, and its characteristics imply the absence of discrete atoms

The support of the National Science Foundation during the early part of this study is gratefully acknowledged.

APPENDIX

Strong coupling is assumed because the assumptions of the adiabatic approximation are not fulfilled. The acoustical velocity is 240 meters per second as compared to the velocity of helium atoms in the $J=2$ rotational state in a 3.7 Å-diameter orbit of 170 meters per second. These velocities do not differ sufficiently to justify weak coupling.

In the present theory, strong coupling is achieved by considering an inhomogeneous space of four dimensions in which the three equivalent axes are Debye coordinates but the fourth is associated with rotation of atoms in pairs. A four-dimensional spheroid has the formula:

$$\frac{W^2}{a^2} + \frac{X^2}{b^2} + \frac{Y^2}{b^2} + \frac{Z^2}{b^2} = 1$$

The volume of the spheroid is

$$v = \frac{\pi^2}{2} b^3 a$$

and the differential volume with respect to b is

$$dv = \frac{3}{2} \pi^2 a b^2 db$$

Debye theory gives $b = 2\nu/c$ where c is the velocity of sound, and ν is the frequency of the acoustical mode. The number of cells available becomes

$$dZ = \frac{12\pi^2 a}{c^3} \nu^2 d\nu$$

and the internal energy per mole is

$$dU = \frac{Vh\nu dZ}{e^{h\nu/kT} - 1}$$

where V is the molecular volume, h, Planck's constant, k, Boltzman's constant, and T the absolute temperature.

The expression for a is taken from the energy of rotation of a diatomic gas and is the fraction of the molecules in rotation multiplied by an integer presently assumed to be $J\,(J+1)$

$$a = \frac{\sum_{J=2,4,\cdot} (2J+1)e^{-J(J+1)B/kT}J(J+1)}{\sum_{J=0,2,4,\cdot} (2J+1)e^{-J(J+1)B/kT}}$$

B is the rotational constant $h^2/8\pi^2 I$ with I the moment of inertia J is the rotational quantum number which for homonuclear atoms with zero nuclear spin is confined to $0, 2, 4, \ldots$. In nearly every instance, contributions from $J=4$ and larger are trivial, so a may be approximated by

$$a = \frac{30e^{-6B/kT}}{1+5e^{-6B/kT}}$$

The low temperature approximation of Debye may be assumed since T_λ is one tenth of the Debye temperature The internal energy and heat capacity of the helion anomaly then become

$$U = \frac{24\pi^5 V k^4 T^4}{c^3 h^3}\,\frac{e^{-6B/kT}}{1+5e^{-6B/kT}}$$

$$C = \frac{48\pi^5 V k^4 T^3}{c^3 h^3}\left[\frac{2e^{-6B/kT}}{1+5e^{-6B/kT}} + \frac{3Be^{-6B/kT}}{kT(1+5e^{-6B/kT})^2}\right]$$

To this helion heat capacity must be added the Debye (phonon) contribution which will predominate below 0 7°K but which is not significant above 1°K.

Empirical constants entering the expression are the molecular volume (55 cm³/8 gm mole), the velocity of sound (237 m/sec) and the nearest neighbor distance (3.7 Å) which enters the moment of inertia of the rotor

The normal component is accurately correlated with the

internal energy. Landau derived an expression for the density of the normal component $\rho_n = 4U/3c^2$ where c is the velocity of propagation of the excitation: for phonons, the velocity of first sound; for helions, the velocity of second sound (18.5 m/sec).

The number of helions may be calculated

$$dN = \frac{V}{e^{h\nu/kT} - 1}\, dZ$$

If the low temperature approximation is used, the integrated expression becomes

$$N = \frac{2.404(360)Vk^3T^3}{c^3h^3}\, \frac{e^{-6B/kT}}{1 + 5e^{-6B/kT}}$$

If N is Avogadro's number, a temperature of 2.7°K is obtained, analogous to the Debye temperature (20°K). This is identified with the lambda point. In a solution of helium-3, in helium-4 only He^4–He^4 pairs can show the postulated quantized rotation. If pairs occur statistically, the concentration will be proportional to the square of the mole fraction $N_{44} = X_4^2 N$. The calculation of N_{44} gives a curve of T_λ as a function of mole fraction. (In this calculation $J = 4$ may be detected for nearly pure helium-4.)

The density of the liquid can be calculated by assuming a perfect solution of two components, the Debye liquid and the roton or helion component, differing in intrinsic density by 1.2 per cent but having the same expansion coefficient ($\alpha = 7.5 + 10^{-4}\, T^3$).

To obtain the dispersion relation, the expression for the internal energy of this theory is equated to the classical expression

$$\frac{4\pi V}{h^3} \int \epsilon e^{-\epsilon/kT} p^2 dp = \frac{4\pi V}{h^3} \int \frac{(1 + 3\pi a)h\nu}{e^{h\nu/kT} - 1}\, p^2 dp$$

Since the limits of integration are identical, the integrands may be equated in the approximation of Boltzman statistics,

$$\epsilon e^{-\epsilon/kT} = (1 + 3\pi a)\, h\nu e^{-h\nu/kT}$$

Debye's relative coordinate $x = h\nu/kT$ is introduced together with an additional adjustable constant alpha

$$\epsilon e^{-\epsilon\alpha x/h} = \left(1 + \frac{90\pi e^{-6Bx/h}}{1 + 5e^{-6Bx/h}}\right) h\nu e^{-\alpha x}$$

If x is assumed constant, the frequency ν is formally analogous to the temperature in the earlier expressions. Since frequency is proportional to momentum, momentum states will contribute in sequence to the energy. Again these momentum states may be counted to Avogadro's number so that the lambda temperature should have an equivalent momentum The minimum of the dispersion curve at 1.95 $Å^{-1}$ is therefore identified with the lambda point at 2 18°K. This permits the evaluation of the constant x as 16 5, a value which gives a good fit where the curve departs from the phonon line. The theoretical fit in Figure 1 is achieved with $x = 18$ and $\alpha = 0.3$ No *a priori* justification can be given for this value of alpha at this time

REFERENCES

1. K MENDELSSOHN, *Cryophysics*, Interscience, New York (1960)
2. C T LANE, *Superfluid Physics*, McGraw-Hill, New York (1962)
3. K R ATKINS, *Liquid Helium*, Cambridge Univ. Press, London (1959)
4. L D LANDAU, *J Phys*, *Moscow*, 5, 71 (1941), *11*, 91 (1947)
5. R P FEYNMAN, *Phys Rev*, *91*, 1291, 1301 (1953), *94*, 262 (1954)
   R P FEYNMAN and M COHEN, *Phys Rev*, *102*, 1189 (1956)
6. J DE BOER, *Progress in Low Temperature Physics*, 2, 1 (1957)
7. H. N. V. TEMPERLEY, *Proc Phys Soc*, *London*, *68A*, 1136 (1955)

3. HUMANISTIC BIOLOGY

By René Dubos
The Rockefeller University

Man's Nature and Man's History

The conjunction of the two words "humanistic" and "biology" will probably seem artificial because very few scientists, and even fewer humanists, really believe that biological knowledge has relevance to the traits that account for the humanness of man Admittedly, biological determinants do not seem at first sight to play a significant role in the manifestations of life which are most characteristically human, for example, ecstasy, logic, or simply the experience of happiness and dispair Religious and ethical doctrines, philosophy, linguistics, literature, the arts, are part of the humanities because their problems obviously relate to the social and cultural history of man, but their connections with the biological attributes of *Homo sapiens* are not so readily apparent "Man has no nature, what he has is history," wrote Ortega y Gasset

While it is obvious that man is the product of his social and cultural history, it is equally certain, on the other hand, that everything he does is conditioned by his biological attributes The performance of each human being, and of each human group, reflects biological necessities and propensities inherited from the evolutionary and experiential past. Human decisions create social and cultural history, but the raw materials of this edifice are derived from man's biological history

It is always dangerous, of course, to use biological terminology and concepts in discussing human affairs. Half a century ago, for example, the illustrious biologist Elie Metchnikoff published under the title *Etudes sur la Nature Humaine* a book in which he attempted to explain the properties of the human flesh and the manifestations of the human spirit in the light of the biological knowledge of his time. His English translator, the zoologist P. Chalmers Mitchell, found it difficult to convey the full meaning of the expression *Nature Humaine* in Metchnikoff's title. As he wrote in the preface to his translation, "*Human nature* is not an exact equivalent of *la nature humaine,* for the latter phrase has a more complete significance, and definitely implies, not only the mental qualities of man, but his bodily framework, with its inherited and acquired anatomical structure and body functions." Mitchell thought that "The Nature of Man" was a more accurate rendering of Metchnikoff's title than "Human Nature." In 1930, the American biologist H. S. Jennings also recognized implicitly the limitations of the phrase "human nature" when he qualified it as *The Biological Basis of Human Nature* in the title of a book which he devoted to "those aspects of experimental biology that are of most interest in considering the problems of human personality and society."

Throughout the present essay, I shall use the phrase "man's nature" instead of "human nature" to emphasize my conviction that the somewhat limited meaning implied in the English usage of human nature does not convey the depth and richness of the knowledge which biological sciences could bring to bear on cultural and social history. I shall attempt to show that the psychological and ethical attributes of man, and the preoccupations that constitute his humanness, are inseparable from the physiological needs and urges which biological experience has inscribed in his flesh and bones. My theme is that better knowledge of man's nature

would enlarge and deepen the understanding of man's history

The doctrine of evolution has made it obvious that living organisms cannot be understood except in the light of their past. Surprisingly enough, however, the science of human biology has been little influenced by the historical outlook. The knowledge of man has progressed far, of course, since Descartes and his followers tried to base it on the study of the mechanical models with which they were familiar, but its most spectacular advances have been along the road first opened in the seventeenth century. There is an almost universal tendency to identify the "science" of man's nature with the physicochemical description of the body's structures and mechanisms, and with the laws which govern the transmission of its hereditary characteristics Yet human life clearly constitutes an experience far more complex than the phenomena which are encompassed by this limited approach, its dynamic processes constantly alter the physical and mental structure of man as he responds to the challenges of the natural environment and of the ways of life that he himself creates. The almost complete irrelevance of present-day biology to the humanities comes largely from the fact that it emphasizes the mechanical aspects of man's nature rather than his experiences, it is concerned more with his being than with his responding and becoming

Increasingly during recent decades, the exact biological sciences have been focused on the phenomena that are common to all mammalian species and indeed to all living forms. This trend away from the special attributes which particularize human beings makes scientific biology appear even further remote from humanistic preoccupations The kind of knowledge to which it leads throws very little light on the problems which are the primary concern of humanists, namely the human experiences in a particular culture. Yet, while they appear at first sight so coldly detached from living

man, the findings of orthodox biological sciences have nevertheless profoundly influenced some of the largest philosophical expressions of modern humanism.

Man in the Great Chain of Being

The doctrine of evolution has influenced all aspects of modern culture by providing biological evidence for the social concept of historical change. Surprising as it may seem to us, it is only during modern times that the myth of eternal return has been displaced in the Western mind by the concept of progress, namely the belief that the universe, and especially the world of men, is constantly moving toward states differing not only from those of the present but also from anything in the past. Whatever its origins, the doctrine of progress did not become part of collective consciousness until the theory of biological evolution provided a scientific model for it.

Most enlightened persons now accept as a fact that everything in the cosmos—from heavenly bodies to human beings—has developed and continues to develop through evolutionary processes. The great religions of the West have come to accept a historical view of creation. Evolutionary concepts are applied also to social institutions and to the arts. Indeed, most political parties, as well as schools of theology, sociology, history, or arts, teach these concepts and make them the basis of their doctrines. Thus, theoretical biology now pervades all of Western culture indirectly through the concept of progressive historical change.

One of the most relevant contributions of biological science to sociology has been to provide a scientific basis for the ancient ethical and religious doctrines of the brotherhood of man, by demonstrating that all human beings belong to the same species. Comparative biology has revealed, furthermore, that man is linked to all living organisms through

a common line of descent, and shares with them many characteristics of physicochemical constitution and of biological organization, the philosophical concept of the "great chain of being" can thus be restated now in the form of a scientific generalization. Even St Francis' love of his brothers the beasts and the Hindoo reverence for life can be regarded as the poetical, ethical, and philosophical expressions of the biological law that all living forms exhibit a deep underlying unity even though they are so strikingly diverse.

Paradoxically, the very success of comparative biology and evolutionary doctrines in relating man to the rest of creation may have retarded the growth of knowledge concerning man himself Since all living forms have so many characteristics in common, biologists and even medical scientists naturally tend to focus their investigative efforts on organisms which are simpler than man, and therefore easier to manipulate in the laboratory This tendency is based on the widespread, though unproven assumption, that understanding of man will eventually emerge from detailed knowledge of the elementary structures and functions which occur in all living things One of the deplorable consequences of this attitude is the common belief that the only fields of biology deserving to be called "fundamental" are those that deal with the simplest manifestations of life, and preferably with *lifeless* reactions and structures derived from living things! Yet it is certain that such a limited approach is not sufficient to create a science of life, let alone of man

Vague though they are, words such as "mind" and "emotion" nevertheless symbolize essential aspects of human life which cannot be related to isolated morphological structures, or formulated in physicochemical terms Thus, some of the most interesting aspects of human life, and certainly the most influential, do not come within the purview of what is at present called "fundamental" biology. I hasten to

in other words, that the increase in size of the human brain occurred simultaneously with the first phases in the unfolding of human culture. Man and his culture evolved, simultaneously, as it were, through a complex series of feedback processes

It seems a reasonable kind of science fiction to imagine that the first subhuman creature who used a tool thereby opened an evolutionary channel in which greater ability to use tools provided a selective advantage. Through analogous evolutionary mechanisms, somatic changes and organization of reaction patterns followed upon the development of family structure and of practices for hunting in groups Soon after, perhaps, the primitive forms of art, of religion, and even of "science" also played their part in affecting the development of neural processes and their integration. New reaction patterns thus became progressively molded on the ways of life, as the brain enlarged. And reciprocally, the ways of life evolved as the brain and its functions became better fitted to them and more complex

The "higher" functions of the human brain probably result therefore from progressive structural–functional transformations giving rise to a system which permits the physiological and behavioral adaptation of man to his own culture By necessity, of course, this adaptation cannot be perfect, since culture is continuously evolving What is almost certain, however, is that the various components of human culture are now required not only for the survival of man but also for his existential realization. Man created himself even as he created his culture and thereby he became dependent on it

Another fact relating biological studies to humanistic problems is that the higher the position of an animal on the phylogenetic ladder, the more unpredictable is its behavior with regard to environmental stimuli Up to the lower

mammals, the emergence of relative independence from external influences can be correlated with the appearance of novel neural mechanisms, but no such anatomical substratum is yet known to account for the high degree of freedom exhibited by the higher mammals. Irrespective of explanation, however, it can hardly be doubted that man is the most evolved organism with regard to the degree of his independence from the direct effects of the physicochemical environment. One of the most profitable approaches to the definition of *Homo sapiens* might therefore be to describe the mechanisms through which his evolutionary ancestors have progressively increased their experiential independence, thus creating his biological identity. But I shall limit myself here to a few remarks concerning man as he exists today, or rather concerning his reactions with, and his responses to, the forces which impinge on him.

I have purposefully differentiated reaction from response because those two words symbolize in my view the two extreme ends of the wide spectrum of interplay of man and his environment. From one end of the spectrum, man appears as an ordinary physiocochemical machine, complex of course but nevertheless reacting with environmental forces according to the same laws that govern inanimate matter. From the other end, man is seen as a creature which is rarely a passive component in the reacting system; the most characteristic aspect of his behavior is the fact that he responds actively and often creatively. Man is able to shut out some of the stimuli to which he is exposed; he modifies others through symbolic and sociocultural mechanisms; most importantly, he can use the effects of stimuli to his own selected ends. All degrees of variation exist between the passive reactions with the environment and the creative responses through which the personality asserts itself. Man is the more human the better he is able to convert passive reactions into creative responses.

The Divine Madness

The performance of any living organism in any given situation is conditioned of course by environmental forces. But its characteristics are determined by the potentialities and the limitations which the organism has acquired and retained from its evolutionary and experiential past. Admittedly, man often behaves as if he were completely independent of his biological history. However, while his outward behavior reveals such a large degree of freedom, physiological reactions elicited in him by environmental and sociocultural forces appear to be very similar to those manifested by his Paleolithic ancestors His ancient needs and urges persist even when their overt manifestations are so masked or distorted that he himself is not aware of their existence. The survival of the Paleolithic past in modern man accounts for many puzzling aspects of his responses to the total environment—for the pathological as well as the expressive and creative aspects of his behavior

The extraordinary degree to which the physiological processes of human life are still linked to cosmic rhythms provides a striking illustration of the persistence of traits having their origin in man's evolutionary past Modern man is wont to boast that he can control his external environment; he can indeed illuminate his rooms at night, heat them during the winter, and cool them during the summer, he can secure an ample and varied supply of food throughout the year But even when he elects to follow unchangeable ways of life in an environment which appears uniform, all the functions of his body continue to fluctuate according to certain rhythms linked to the movements of the earth and of the moon with respect to each other and to the sun His hormonal activities, in particular, exhibit marked diurnal and seasonal rhy
and probably other rhythms also are linked to those
cosmos.

All aspects of behavior are affected by physiological processes. Man's responses to any situation are different in the morning from what they are at night, and different in the spring from what they are in the autumn. The writers of Western stories are on a sound biological basis when they recount that the Indians always attacked at dawn, because the spirits of the white man were then at a low ebb. The wild imaginings of the night, and the fears which they engender, are indirectly the effects of the earth movements, because the human organism readily escapes from the control of reason under the influence of the physiological changes associated with darkness. The lunar cycles are also reflected in the physiology and behavior of animals and probably therefore of man. It would not be surprising if the moon worshippers as well as the "lunatics" were really affected—as the words suggest—by lunar forces to which all of us are also sensitive.

Seasonal changes, in any case, certainly affect most living things, including man, even when the temperature and illumination are artificially maintained at a constant level. In the most mechanized, treeless, and birdless city, just as in the hills of Arcadia long ago, men and women perceive in their senses and reveal by their behavior that the exuberance of springtime and the despondency of late fall have origins more subtle than the mere change in temperature. It is for good biological reasons that carnival and Mardi Gras are celebrated when the sap starts running up the trees, and that men commemorate their dead in late fall when nature is dying. Thus, modern man in his sheltered environment continues to be under the influence of cosmic forces much as he was when he lived naked in direct contact with nature. Similarly, he continues to react physiologically to the presence of strange living things, and especially of human competitors, as if he were in danger of being physically attacked by them. The fight-or-flight response, with all its deep physiological

accompaniments, is a biological carry-over from the time when the survival of primitive man encountering a wild animal or a human stranger depended upon his ability to mobilize the body mechanisms which enabled him to engage in physical struggle or to flee.

Man, on the other hand, evolved as a social animal, and he can neither fully develop, nor function normally, except in association with other human beings. All social stimuli express themselves in the forms of phenomena which in their turn condition the response to the life situations which had evoked them Thus, crowding, isolation, challenge of any sort have effects which have their origin in the evolutionary past, and tend to imitate the kind of response that was then favorable for biological success—even when such a response is no longer suitable to the conditions of the modern world Many aspects of human behavior which appear incomprehensible, or even irrational, become meaningful when interpreted as survivals of attributes that were useful when they first appeared during evolutionary development and have persisted because the physical evolution of man came to a relative halt about 150,000 years ago Phenomena ranging all the way from the aberrations of mob psychology to the useless disturbances of metabolism and circulation which occur during verbal conflicts at the office or at a cocktail party are as much the indirect expressions of the distant biological past as they are the direct consequences of the stimuli that were their immediate causes

The urge to control property and to dominate one's peers are also ancient biological traits which can be recognized in the different forms of territoriality and dominance among most if not all animal societies Animal behavior provides prototypes of the lust for political power, independently of any desire for financial or other material rewards, which is so common among men Even the play instinct and certain kinds of aesthetic expression correspond to derivative but

nevertheless important biological needs which exist in one form or another in animal species and which have probably always been part of man's nature.

These biological characteristics and many others that cannot be mentioned here are woven into the very fabric of the human race, and they condition all aspects of human behavior. Unfortunately, they have been grossly neglected by biologists. This neglect is the result of the historical accident that scientific biology has been identified from its very beginning with the concept that the body is a complex but otherwise ordinary machine, and that detailed analysis of its elementary structures and energy mechanisms is the only valid approach to the understanding of the living organism. Such an attitude had discouraged the scientific study of the biological problems that do not lend themselves to the reductionist analytic methods now in vogue among experimental scientists. In particular, it has inhibited the study of the biological phenomena which are the consequence of the organism's evolutionary history—for example, the manifestations of the ancient urges which do not come readily to the surface under the conditions of civilized life, or the effect of the season and the hour of the day on the reactions elicited by physicochemical stimuli and various life situations.

Yet, numerous observations have now established that the oscillations in physiological and mental processes affect all activities of the organism, and even the efficacy of therapeutic procedures. The orthodox reductionist approach is not suitable for the study of these phenomena, because biological clocks, like the mind, disappear when the organism is dissected into its lifeless component parts. In fact, the most important problems of life can be recognized only when the organism responds actively to its environment as an integrated unit. Interestingly enough, space research is at present creating a wave of interest in topics as earthy as the effects on man of the tides, the seasons, and the diurnal cycles. Just

as rockets and satellites are giving new importance to celestial mechanics, so does the prospect of space travel call attention to the potentialities and constraints which are the consequences of biological rhythms.

Philosophers, writers, and artists have always been acutely aware of the immense role played by occult biological processes in human life. In *Phaedrus*, Socrates speaks with passion of the creative forces released in man by the "divine madness." The text of the dialogue makes it clear that the word "madness" as used by Plato refers not to pathological mental states but rather to those deep biological attributes of man's nature which are almost beyond the control of reason and transcend its reach. These attributes remain concealed under the usual circumstances of ordinary life, but they constitute the most powerful sources of inspiration for the scientist as well as the artist. Creativity depends in part on the ability to hear "the voice of the deep" and to tap resources from regions of man's nature which have not yet been explored.

Nietzsche was referring to innate forces analogous to Socrates' divine madness when he wrote in *The Birth of Tragedy* that the Dionysian inspiration is a necessary complement of the Apollinian order. In fact, as shown by Dodds in *The Greeks and the Irrational*, ancient civilizations were aware of the existence of these powerful biological needs of man's nature which are not clearly perceived and thus appear irrational. They symbolized the occult passions—the divine madness—by a ferocious bull struggling against reason.

Empirically, all over the world, social practices have been developed to let these occult forces manifest themselves under somewhat controlled conditions. The Dionysian celebrations, the Eleusinian mysteries, and many other myths and rituals served as release mechanisms for biological urges which could not find an otherwise acceptable expression in the rational aspects of Greek life, even Socrates participated in the Corybantic rites. Needless to say, such ancient tradi-

tions still persist even in the most advanced countries of the Western world, though often in a distorted form. Even when man has become an urbane city dweller, the Paleolithic bull which survives in his inner self still paws the earth whenever a threatening gesture is made on the social scene. The passions depicted by classical tragedies have their roots deep in the Paleolithic past.

Experiential Innocence

All human beings have fundamentally the same anatomical structure, function through the same chemical activities, exhibit the same physiological manifestations, and even possess the same occult biological needs; yet no two human beings are alike. Clearly, knowledge of the attributes which are common to mankind as a whole is not sufficient to account for the manner in which each individual person behaves as he does, develops his own peculiarities, in brief becomes different from all other human beings. Since man's sense of discreteness is among his most cherished and most pronounced characteristics, the failure of theoretical biology to emphasize the uniqueness of individual human beings contributes to its lack of influence on the humanities.

Individual persons differ, of course, by reason of the fact that they do not have the same genetic makeup; only identical twins are alike genetically. But at least equally important is the fact that the individual characteristics of human beings are constantly being shaped and modified by environmental factors which endlessly vary with time, differ from one place to the other, and therefore are never the same for two different persons. Recent studies have confirmed the ancient knowledge that many characteristics of the adult result from the effects of so-called "early influences," namely of those environmental factors which impinge on the organism during early life, while it is still developing. Such formative effects

can take place even *in utero.* For example, even though the Dionne quintuplets were genetically identical, they could be differentiated shortly after birth, probably because their different positions during intrauterine life had differentially affected their development. Prenatal and early postnatal influences can affect almost every trait—from nutritional needs and rates of growth, to learning ability and emotional attitudes *Moreover, the effects of early influences are so deeply* rooted in the biological structure of the person involved that they often and perhaps always persist throughout his whole life span

Environmental influences shape personality through two different groups of mechanisms. On the one hand, they determine certain patterns of response which can affect all manifestations of behavior Physiologists, psychologists, psychiatrists, and novelists have described, each in his own way, a seemingly endless variety of conditioned responses ranging from the imprinting of ducks by early associations with a foreign object or the salivation of dogs at the sound of a bell, to the pathological effects of the Freudian complexes or the remembrance of things past evoked by a madeleine dipped into a cup of tea The widespread belief in the existence of an all embracing and lasting biological memory is symbolized by Tennyson's statement in *Ulysses*, "I am a part of all that I have met "

On the other hand, environmental influences contribute also to the shaping of the personality by interfering with the acquisition of new experiences Ideally, man should remain receptive to new stimuli, new events, and new situations in order to continue developing mentally But in fact, the aptitude to apprehend the external world with freshness of perception commonly becomes saturated as the mind and senses become conditioned by repeated experiences.

The environmental influences that are ubiquitous in a given geographical, national, or social group naturally tend to bring

out many characteristics which are common to all members of the group. For this reason, there is much truth in Emerson's statement that "We resemble our contemporaries: even more than our progenitors." But environmental influences also affect each person in an individual manner, even when the ways of life appear uniform in a standardized environment. Genetic uniqueness makes for differences in response and consequently in mental and physical development; furthermore, each one of us lives in a private world of his own.

Human beings thus perceive the world, and respond to it, not through the whole spectrum of their potentialities but only through the areas of this spectrum which have been made functional by environmental influences, especially the early ones, and which are not blocked by inhibitory mechaisms. I have intentionally formulated this problem in very general and consequently vague terms, because the word potentiality is meant here to denote the whole range of the organism's genetic endowment, in both its physical and mental expressions. The life experiences determine what parts of this endowment emerge in the form of functional attributes.

In obscure ways, the kind of creation associated with the word genius must be related to the developmental processes that determine the manner in which the mind responds to external stimuli. Whether it manifests itself by revealing heretofore unrecognized aspects of reality, or by making new patterns out of facts already known, creativity often involves the ability to contemplate the world with a holistic and unconditioned attitude. Complete receptivity, however, is the prerogative of childhood, and of the few privileged adults who have retained or recaptured the directness of perception and experiential innocence which enable them to perceive "things as they are." Hence, the deep biological truth of Baudelaire's arresting and visionary image, "Le génie, c'est l'enfance retrouvée." Genius is childhood recaptured.

Unfortunately, it appears that the knowledge acquired

through the practice of daily life or by systematic learning must often be paid by loss of ability for truly original creation. Brancusi's statement that "when we are no longer young, we are already dead" expresses the cruel but inescapable truth that the mature stages of life are often encumbered with traditional and conventioanl attitudes which interfere with the receptivity for new experiences. The fact that the human *personality tends to "set" with age suggests some functional* "rigidification" of the potential ability for continual progressive changes in the engrams stored in the brain—as if there were sharp limitations to the experiential life span Thus, one of the great problems of biology is to determine whether the effects of early influences are truly irreversible, as ordinary experience seems to show, or whether they can be erased partially at least, as a few animal experiments suggest

Electrophysiological studies have revealed that the activity of neural processes in the brain is continuous, and that the effect of stimuli is to give form to the activity going on, rather than to arouse inactive tissue These findings, as well as the knowledge that sensory deprivation commonly results in a *transient disintegration of the personality, seem to suggest* that ways may be found to prevent or retard the "setting" of personality so that man's body does not outlive the better of his mental faculties In practice, however, the only evidence suggesting that the mind can be reshaped has come, tragically, from brainwashing or confessions of political guilt, and, more hopefully, from certain forms of religious conversions

The loss of experiential innocence during childhood is of course inevitable and is indeed a *sine qua non* of mental and emotional growth But what matters is the kind of experience through which innocence is lost, because this determines in large measure the shaping of the personality The ancient dream of the Fountain of Youth might acquire a new and richer meaning if an acceptable technique could be developed to re-establish a state of receptivity in the fully developed

and conditioned organism. Recapturing childhood, in Baudelaire's sense, implies reacquiring the ability for direct apprehension of the external world. It would be the surest approach to a true enlargement of human life.

Man Creates Himself

The fundamental biological nature of man has not changed since late Paleolithic times. The same arrangement of 20,000 pairs of genes still controls his physical development and his physiological reactions; the implements he made during the Stone Age still fit his hands; the ancient drives that shaped his cultural evolution are still operative; the rituals and symbolic representations which he performed in the Paleolithic caves are still meaningful to him. But while *Homo sapiens* has remained essentially the same, the manifestations of his life, and the structure of his societies, are endlessly changing, never repeating themselves identically. The permanency comes from the nature of the raw materials out of which human beings are made, the change from the creative responses which man makes to the challenges of his total environment. To live is to function, which means to respond.

Biological science has been immensely successful in describing the structures and mechanisms which constitute living *things,* and through which they operate. But it has contributed much less to the knowledge of man as a functioning organism. Yet the human condition cannot be dealt with scientifically unless a systematic effort is made to describe and analyze the pattern of responses that man makes to all the stimuli which impinge on him. For it is precisely this pattern that defines the human condition. In my judgment, such knowledge could be acquired if biologists elected to devote to the study of the living experience as much skill and energy as they have devoted to the description of the body machine. Fortunately, they can undertake this task with the confidence that they

will find in the animal kingdom experimental models for many of the most interesting problems of human life.

The paradox is that while man unquestionably occupies a unique place in creation, his responses to environmental stimuli have their counterparts in the life of one or another animal species Man's physiological urges, including the need to play, are common aspects of animal life in nature; most of his social activities and organizations are also represented in the different types of animal communities; even the ability to express attitudes and desires through symbolic sounds, postures, objects, and other representations is widespread among animals Furthermore, it has been clear ever since Kropotkin published his book, *Mutual Aid, A Factor in Evolution*, that social attitudes arising from biological necessities can evolve into ethical principles In brief, it will probably be possible to find somewhere in the animal world experimental models suitable for the study of many aspects of human life.

While models are useful and indeed essential for the scientific analysis of particular problems, they cannot provide a complete knowledge of man Models never truly represent nature. This limitation is not peculiar to the knowledge of man, or of other living organisms; it applies also to inanimate nature as well On this score, I shall content myself with quoting statements made by Eugene P Wigner in Stockholm when he accepted the Nobel Prize in Physics in 1963

> Physics does not endeavor to explain nature In fact, the great success of physics is due to a restriction of its objectives: it endeavors to explain the regularities in the behavior of objects. This renunciation of the broader aim, and the specification of the domain for which an explanation can be sought, now appears to us an obvious necessity. In fact, the specification of the explainable may have been the greatest discovery of physics so far
>
> The regularities in the phenomena which physical sci-

and conditioned organism. Recapturing childhood, in Baudelaire's sense, implies reacquiring the ability for direct apprehension of the external world. It would be the surest approach to a true enlargement of human life.

Man Creates Himself

The fundamental biological nature of man has not changed since late Paleolithic times. The same arrangement of 20,000 pairs of genes still controls his physical development and his physiological reactions; the implements he made during the Stone Age still fit his hands; the ancient drives that shaped his cultural evolution are still operative; the rituals and symbolic representations which he performed in the Paleolithic caves are still meaningful to him. But while *Homo sapiens* has remained essentially the same, the manifestations of his life, and the structure of his societies, are endlessly changing, never repeating themselves identically. The permanency comes from the nature of the raw materials out of which human beings are made, the change from the creative responses which man makes to the challenges of his total environment. To live is to function, which means to respond.

Biological science has been immensely successful in describing the structures and mechanisms which constitute living *things*, and through which they operate. But it has contributed much less to the knowledge of man as a functioning organism. Yet the human condition cannot be dealt with scientifically unless a systematic effort is made to describe and analyze the pattern of responses that man makes to all the stimuli which impinge on him. For it is precisely this pattern that defines the human condition. In my judgment, such knowledge could be acquired if biologists elected to devote to the study of the living experience as much skill and energy as they have devoted to the description of the body machine. Fortunately, they can undertake this task with the confidence that they

will find in the animal kingdom experimental models for many of the most interesting problems of human life.

The paradox is that while man unquestionably occupies a unique place in creation, his responses to environmental stimuli have their counterparts in the life of one or another animal species Man's physiological urges, including the need to play, are common aspects of animal life in nature; most of his social activities and organizations are also represented in the different types of animal communities; even the ability to express attitudes and desires through symbolic sounds, postures, objects, and other representations is widespread among animals Furthermore, it has been clear ever since Kropotkin published his book, *Mutual Aid, A Factor in Evolution,* that social attitudes arising from biological necessities can evolve into ethical principles In brief, it will probably be possible to find somewhere in the animal world experimental models suitable for the study of many aspects of human life.

While models are useful and indeed essential for the scientific analysis of particular problems, they cannot provide a complete knowledge of man Models never truly represent nature. This limitation is not peculiar to the knowledge of man, or of other living organisms; it applies also to inanimate nature as well. On this score, I shall content myself with quoting statements made by Eugene P. Wigner in Stockholm when he accepted the Nobel Prize in Physics in 1963:

> Physics does not endeavor to explain nature In fact, the great success of physics is due to a restriction of its objectives: it endeavors to explain the regularities in the behavior of objects. This renunciation of the broader aim, and the specification of the domain for which an explanation can be sought, now appears to us an obvious necessity. In fact, the specification of the explainable may have been the greatest discovery of physics so far
>
> The regularities in the phenomena which physical sci-

> ence endeavors to uncover are called the laws of nature. The name is actually very appropriate. Just as legal laws regulate actions and behavior under certain conditions but do not try to regulate all actions and behavior, the laws of physics also determine the behavior of its objects of interest only under certain well-defined conditions but leave much freedom otherwise.

The word freedom as used by the physicist has a meaning which is very different, of course, from the one it has when applied to human affairs. Nevertheless, even though the analogy is only formal, I shall use it in an attempt to state my faith concerning the kind of contribution that scientific biology can make to the humanities. First, however, I must emphasize that the biological information discussed in the preceding pages was not intended to give an account of the living, throbbing man whom the humanist tries to apprehend and the artist to express. Its role was only to describe the raw materials out of which man creates himself, through a continuous series of personal decisions. A few remarks concerning the plastic arts will serve to illustrate my view of the manner in which man's ability to choose and decide determines and limits the relevance of biological knowledge to the humanities.

The paintings, statues, engravings, and other artifacts found in Paleolithic sites leave no doubt that the faculty for artistic expression is very ancient; indeed, it does not seem to have improved with regard either to acuity of perception or to skill in representation over the past 20,000 years. There is consequently every reason to believe that the ability to perceive and to represent corresponds to deeply seated physiological attributes of man's nature. This aesthetic faculty is probably analogous to that which makes animals perform movements or build nests which have an intrinsic harmony. Certain colors clash or certain elements in a drawing appear to be out of proportion, when the experience of them con-

flicts with some built-in set of relationships in our own organs. In contrast, we would not be pleased with the way certain things look unless our organs and senses were so constituted as to be in harmony with the proportions and rhythms associated with these things. Whether the manner in which the senses or organs respond has a purely genetic basis or is the product of conditioning influences, is irrelevant here. The point I want to convey is that aesthetic consciousness depends on innate faculties which are biological in essence and which do not seem to have progressed since Paleolithic times.

However, the ability to perceive and to represent is not sufficient by itself to create works of art; other factors are involved which fall outside the realm of the biological sciences. While human beings respond to their environment through their biological attributes, they do not react passively as if they were but mechanical intermediaries in stimulus–response couplets. The artist's response is not mechanical, nor even simply motivated by the need to cope with the environment. It constitutes rather an expressive behavior in which the artist uses the environment for the purpose of self-actualization

The act of artistic creation thus provides a convenient example to illustrate the role of human choices in willing a certain course of action among the possibilities of dealing with external nature through the needs, drives, and urges which are inherent in man's nature Similarly, all other aspects of human life present opportunities for active intervention on the part of man—the creature who can choose, eliminate, assemble, decide, and thereby create

As is well known, man has now the technical means to transform his life and himself by manipulating his environment as well as his physiological and mental processes, soon he may even be able to alter somewhat his genetic makeup The powers of action generated by scientific advances are so great that the classical discussions on the ideals of the good life now take on very practical meaning. Mankind—that is to say we—

shall find ourselves drifting aimlessly toward a state incompatible with the maintenance of the humanistic values from which we derive our uniqueness, unless we formulate goals worthy of the human condition and are willing to take a stand at the critical time. This kind of freedom is the final criterium of humanness. In the words of Paul Tillich, "Man becomes truly human only at the moment of decision."

Values and goals naturally involve choices and decisions. But the more deeply human life is influenced by technology, the more essential it becomes that these choices and decisions be made in the light of the proper kind of biological knowledge. Man's nature constitutes such a highly integrated system that it cannot be altered safely except within rather narrow limits. Unconsciously, the writers of science fiction acknowledge this fact when they make the human beings who function in an automated world or who travel in spacecraft continue to behave as if they were in a Boy Scout camp or in lovers' lane.

Since radical changes in human life are excluded by biological limitations, imaginings based on limited concepts of the body machine, or on the hope of technological breakthroughs, are of little use in defining what the ideal man should be. The glory of the coming age must be conceived within the framework of man's nature—of his biological limitations as well as his potentialities.

Conclusions

In the very process of responding to environmental stimuli, each individual human being creates his physical and mental personality from the biological attributes which are shared by all men. Human societies and cultures emerged from the progressive integration of these responses.

While the outward manifestations of behavior are governed by the values and rules of each social group, inwardly man

continues to react to his environment in much the same way that he did in the distant past, through the very same physiological mechanisms which he has retained from his evolutionary development. These ancient biological attributes express themselves in the form of reactions which are often unsuited to the modern world and are therefore the cause of difficulties; but, on the other hand, they also constitute a source of inspiration for creative endeavors.

Individual human beings differ not only in their genetic endowment but also by reason of the stimuli to which they have been exposed in early life, both prenatal and postnatal—which give them an experiential uniqueness These early influences shape the physical and mental characteristics by stimulating into activity certain parts of the genetic endowment, and also by inhibiting the expression of others The extent to which a person has retained the ability to apprehend the external world in an unconditioned manner, and to respond to it in his own way, probably plays a large part in his creativeness.

The role of the biologist is to study the raw materials of man's nature and the mechanisms through which each person uses them to create his own experiential individuality This role is becoming of increasing importance as human life becomes more deeply influenced by technology and therefore more remote from man's evolutionary experience

By adding to the knowledge of man's biological nature, science helps the humanist better to understand the human condition and to define the good life Unfortunately, while biological sciences have been immensely successful in describing the elementary structures and processes of the body machine, they have tended to neglect the study of living as experience. Indeed, it is commonly stated that biology has lost contact with the humanities because it has become too "scientific" and as a consequence no longer deals with the problems peculiar to the humanness of man. There is no doubt, of

course, about the loss of contact, but the explanation of the difficulty, in my judgment, is that biology is not scientific enough.

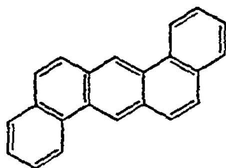
By neglecting the study of a large variety of man's responses, biology is betraying one of the responsibilities of science—namely, the development of objective methods for describing all aspects of reality. Today, as in the past, the most compelling and interesting problems of human life come from the manner in which man reacts passively, and responds creatively, to the challenges of his total environment. Biology will once more become a complementary aspect of the humanities if it accepts the urgent social task to provide knowledge of the raw materials of experience out of which man creates himself.

4. CHEMICAL CARCINOGENS, CARCINOGENESIS, AND CARCINOSTASIS

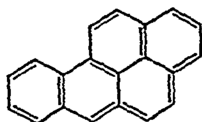
By NORMAN H. CROMWELL
University of Nebraska

Cancer has been defined by Willis[1] as an abnormal mass of tissue, the growth of which exceeds and is uncoordinated with that of normal tissues, which persists in the same excessive manner after the cessation of the stimuli that evoked the change The study of chemical carcinogenesis has proven to be one of the most rewarding and definitive areas of cancer research It was in 1932 that Kennaway[2] and his colleagues in England first demonstrated that cancer could be induced in experimental animals by a pure chemical. Many of the suspected environmental cancers of man have since been reproduced in animals

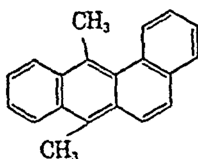
In recent years there has been a renewal and intensification of interest in the causes of cancer with the hope that a better understanding of the processes known as carcinogenesis will eventually lead to the control of this scourge of mankind. Experimental cancer may be induced in three reproducible ways by chemicals, viruses, and ionizing radiation In this brief account an attempt is made to interrelate some contemporary facts and concepts concerned with the carcinogenic (cancer-forming) and carcinostatic (cancer-arresting) actions of certain organic chemicals, with special emphasis on the polycyclic hydrocarbon and heterocyclic compounds



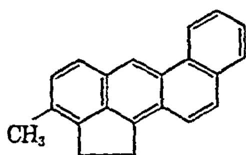
dibenz[a,h]anthracene



benzo[a]pyrene



7,12-dimethylbenz[a]anthracene



3-methylcholanthrene

During the past several decades, a major effort in the general field of experimental oncology has been directed toward the discovery and identification of both carcinogenic and carcinolytic agents. During the latter part of this period, there has been a considerable shift in emphasis from simply synthesizing new agents toward the more basic question of the mechanisms of action for both of these classes of substances. There would seem to be no fundamental reason why, eventually, chemists should not be able to decipher both of these related processes in chemical terms. However, before this can come about, it may well be necessary for us to know more concerning the precise biochemical differences between normal and malignant cells and how these differences control cell division in each instance. Undoubtedly, the causation of cancer and the control or reversal of malignancy will turn out to have closely interrelated mechanistic bases. Both for carcinogenic and carcinostatic agents, there must be a reaction between the agent and some component of the cell for cell growth to be inhibited. In both instances, the question is which component (or components) reacts and is responsible for the biochemical lesion that ultimately leads to the

biological effect. Cancer seems to involve a change in the cells' hereditary mechanisms. A successful carcinostatic agent would be one that would either prevent or reverse this abnormal change in hereditary control of cell growth.

Of the three methods of treating established cancer—surgery, radiation, or chemotherapy—only the first two can be said to effect cures, and these cures represent somewhat over one third of the patients diagnosed as having cancer. If the best possible methods of diagnosis and treatment now available were to be applied, the cure rate might be raised to 50 per cent For the other 50 per cent, it may well be that some form of cancer chemotherapy offers the greatest promise It is of course also possible that we may eventually learn, through the study of carcinogenesis, how to eliminate from our environment, or neutralize the effect of, the major causative factors associated with the initiation and malignant propagation of human cancer.

It has been claimed that, from a practical standpoint, the major problem of finding a cancer cure is to learn how to fight against disseminated cancer cells hidden in remote parts of the body. Some investigators feel that this will require a fight by chemical means If it were not for the widespread dissemination of cancer cells in the patient, he could be cured by the removal of the localized tumor by means of surgery or radiation Thus it has been recognized that there is a great need to study the mechanisms of metastasis in detail, especially in laboratory animals

Much more effort needs to be placed on the study of the mode of action of the drugs now known to have some beneficial effect. Of the twenty to thirty agents now available for the chemotherapy of cancer in man, none provides an enduring cure except for very rare forms of the disease [3] The proper use of these agents, however, makes it possible to hold the disease in check in many cases and to relieve symptoms and prolong useful life. Agents with a much higher

degree of selectivity need to be developed. Even the most effective agents available today are much too toxic to the patient. Drugs with a lower level of toxicity must be found, or new methods of administration must be devised which either protect the patient from the extreme toxic effects of the drugs or potentiate the effectiveness of a lower nontoxic dose of the drug.

Some progress has been made in the more effective administration of known agents. Through use of the surgical techniques known as perfusion and infusion,[4] localized portions of the body are treated with high doses of a cytotoxic agent while the rest of the body is protected by some neutralizing or counteracting chemical. Also, considerable progress in combination therapy has been realized—that is, in the combination of the various mentioned methods, such as an application of chemotherapeutic agents after surgery to attempt to remove free-floating cells,[5] and the combination of cancer chemotherapeutic drugs with radiation to reduce more effectively tumor masses.[6]

It was suggested some time ago that it might be possible to discover special growth stimulants that would cause cancer cells to destroy themselves. Although this happy situation has not yet been realized, the synergetic response of animal tumors and tumors in man to therapy employing either growth factors (corticosteroids) or carcinogens (urethane) in combination with various antitumor agents has been reported by several groups of workers in the United States[6-8] as well as in other countries.[9a] It has been found that several types of compounds including known carcinogenic substances,[8, 9a] or growth-promoting agents,[9b] are capable of sensitizing tumors to the growth-inhibiting action of certain antitumor agents. These observations obviously interrelate carcinogens, carcinogenesis, and carcinostasis.

Among the various drugs which have been found useful in the treatment and palliation of some types of disseminated

neoplastic disease in man are chemical agents that have been classified as cytotoxic agents, antimetabolites, antibiotics, hormones, and alkaloids and miscellaneous compounds for which no basis of action has been suggested. It seems unlikely to expect that the mechanism by which these widely varying substances bring about their biological actions will prove to be precisely the same, unless we mean by "the same" the fact that reactions among the agent, a metabolite or an induced substance, and some component of the cell must be taking place when cell growth is inhibited. One may immediately decide that there are many modes by which this might come about There has been no shortage of theories either to explain carcinogenesis by various types of agents or the carcinolytic (carcinostatic) effect of the better-known antitumor drugs

It must be admitted that theory has not yet played a very successful role in the development of cancer research except as a stimulus to further work. The developments made in recent years have been based almost entirely upon empirical data Many of the present-day theories are mere restatements of experimental or observational findings. The actual mechanism of carcinogenesis is still a matter of speculation. At the moment it seems promising to continue to study it, with the hope that, finally, an adequate picture will emerge which tells us what occurs in tissue and in the cell during abnormal development and growth With this information at hand, we may then hope better to manage the disease in man

The rebirth of interest in carcinogenesis has brought forth a wide variety of active agents of greatly different chemical properties Such widely differing substances as plastic films and metals, and complex organic molecules such as the biological alkylating agents, polycyclic hydrocarbons, and polycyclic heterocyclic compounds are known to initiate cancer Despite the large variety of factors that may be

involved in the initiation of cancerous growth, it appears that the necessary and sufficient alterations that take place with the genetic apparatus of the cells are similar in each case. The particular properties of the cancer cell are propagated to each succeeding generation even after the inducing agent is no longer present. Apparently, the chemical coding ability of the nucleic acid of the cell has been altered.[10]

Theories of Carcinogenesis

Several books[11-14] have been written in the last few years which review the numerous theories and hypotheses advanced to interpret carcinogenesis. In all instances, the authors conclude that these are as yet far from adequate. The most hopeful approaches to a solution of these problems still seem to lie along rather empirical routes such as the further discovery and detailed mechanistic study of agents known to take part in either or both the initiation and promotion stages of carcinogenesis.

One continuing approach to such empirical investigations has been the synthesis and biological study of new members of several series[15-17] of large flat molecules known as polycyclic heterocyclic compounds which have, in common with polycyclic hydrocarbons, the potential biological property of being rather strongly adsorbed on protein surfaces.[18, 19] However, as has been pointed out previously, protein binding is not sufficient to explain carcinogenicity. Many compounds react with proteins but are not carcinogenic (i.e. acylating agents, dinitrofluorobenzene, etc.). The above-named compounds also inactivate enzymes; however, they are not carcinogenic and they do not react with DNA.[20] There are many data indicating that known carcinogenic hydrocarbons bind to proteins,[18, 19] and although it has been widely suggested that these same carcinogenic hydrocarbons bind to nucleic acids, there was little information until recently[21-23] that

this is actually the case.[24] It has been suggested that a possible function of a carcinogenic agent with the ability to bind to proteins may be its removal ("deletion") of a nuclear protein from coordination with DNA, thus allowing the DNA to "run wild" and initiate cancer.[25a] The deletion hypothesis said that tumors resulted from the loss or deletion of an enzyme system responsible for the control of normal growth, possibly because of the combination between the carcinogen and protein. An alternative possibility would be the deletion of some repressor system.[25b]

Clayson[26] has suggested that the two theories which seem of most importance in considering a possible mechanism for chemical carcinogenesis are based on (*i*) somatic mutation, or (*ii*) on the immunological theory of cancer. In spite of the fact that many investigators appear to have assumed that

gens are not carcinogens, and vice versa. Thus it seems unnecessary to identify any portion of the carcinogenic process with mutation. Nevertheless, the change from normal to precancerous tissue during the process of initiation must involve a change in heritable features of the cell.

Green[27] considered malignancy to be the result of the loss of specific antigens by the animal. He suggested that the chemical induction of cancer takes place by two stages: the interaction of the carcinogen with the specific antigens and/or other constituents of the animal tissues to form complexes, an immune response by the host to the tissue–carcinogen complexes accompanied by the elimination of the tissue-specific antigen and the appearance of a neoplastic cell.

It has been generally found that all chemical carcinogens do combine with tissue constituents, i.e. proteins. Kobat[28] has suggested that antibodies react with only certain chemical groupings on the antigenic substance surface.

Some experiments[29] have indicated that the action of

chemical carcinogens in newborn mice does not adequately protect them from later chemical carcinogenesis; i.e. no tolerance is acquired. However, see the section on tumor growth inhibition by polycyclic compounds and Ref. 30.

Indirect Carcinogens

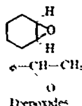
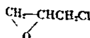
Some effective agents appear not to be direct carcinogens. Boyland[31a, b] has found that the aromatic amines, 2-acetamidofluorene and 2-naphthylamine, are converted *in vivo* to the active compounds, which are probably arylhydroxylamines having chemical properties somewhat similar to the biological alkylating agents that are also carcinogenic. It was found[31b] that 2-naphthylhydroxylamine (50 mg/kg) in arachis oil twice a week for three months, intraperitoneally, in fifteen random inbred strains of albino rats induced six abdominal sarcomas, three abdominal carcinosarcomas of diverse histological types, and one lymphosarcoma. An identical dose of 2-naphthylamine induced sarcomas in $^2/_{14}$ and possibly a salivary gland tumor in $^1/_{14}$. These results agree with the hypothesis that the 2-naphthylamine and some other aromatic amines exert their carcinogenic action after a metabolic conversion to hydroxylamine derivatives.

The arylhydroxylamines react with sulfhydryl compounds to form 5-aminophenyl derivatives (many carcinogenic substances react with sulfhydryl compounds). A few years ago, the Millers and their co-workers[32] at Wisconsin showed that N-hydroxylation is an important reaction which leads to an active intermediate of carcinogenic amino compounds. Ring hydroxylation is the other chief metabolic reaction but seems to be a detoxification route. However, this area is still under active investigation,[33] and the precise details of the mechanism of carcinogenesis by the aromatic amines have not been clearly established. A somewhat complicating factor is the fact that there are sex-linked differences in the metabolism of N-2-fluorenylacetamide by rats.[34]

Boyland[illegible] also suggests that the carcinogenic hydrocarbons, and presumably other polycyclic carcinogens, may be converted by metabolic processes (i.e. conversion to epoxides by oxygen) to carcinogenic derivatives; they may also be oxidized to dicarboxylic acid derivatives leading to combination with proteins and sulfhydryl compounds. These latter metabolic processes with hydrocarbons, however, are probably detoxifications and not involved in the carcinogenic effect which is probably initiated by a charge-complex formation of the purine bases of DNA with either the hydrocarbon itself or an epoxy or peroxy derivative.

Kotin and Falk[illegible] have studied a large number of organic *peroxides and epoxides, and have concluded that they have* a definite carcinogenic and radiometric (DNA chain-breaking) effect (Table 1). This has been pointed out in specific instances previously by several groups of workers, and various members of the class have been known for some time as effective experimental antitumor agents against transplanted tumors.

TABLE 1 EPOXIDES THAT CATALYZE THE DEPOLYMERIZATION OF DNA AND RNA[illegible]

Epoxidated soy oil	
Butyl-9,10-epoxystearate	
	

Cancer-Forming DNA

Recently, Bendich and a group of scientists from the Sloan-Kettering Institute for Cancer Research in New York

collaborated with Stewart and Eddy of the National Institutes of Health to study the nature of the infective and cancer-inducing portion of polyoma virus.[36]

When polyoma virus was inoculated into tissue culture cells of mouse embryo, the virus multiplied and produced the characteristic cytopathic changes and released new virus particles. The virus particles were isolated from the cell culture fluid and their protein coatings stripped off to leave only the bare nucleic acid. This isolated virus nucleic acid was then added to a fresh virus-free mouse embryo cell culture, and again the same cytopathic effects resulted that had been observed with the whole virus. The new virus particles developed in this latter culture were isolated and inoculated into mice or hamsters to produce the disease characteristic of polyoma virus infections. Also, when the isolated nucleic acid itself was injected into the animals, polyoma tumors developed in a small number of them. The isolated nucleic acid was inert to the enzyme RNase but rapidly detoxified by DNase. It was thus established that an infective cancer-producing DNA had been isolated. Since then, the nucleic acid of the Shope virus has been isolated and found to be an infectious DNA.

These findings are of considerable importance and serve as a verification of the general concept that cancer is associated with a permanent change in the cell's genetic center which is controlled by a deoxyribose nucleic acid. The difference in genetic character that leads to variations in the characteristics and functions of different types of cells is based on the differing details of the chemical structures of the DNA's from various species.

Thus, it seems inevitable that cancer cells differ from normal cells in some specific fashion in the structure of the deoxyribonucleic acid genetic center. The malignant powers of these new cancer cells must be directly associated with the particular characteristics of this special DNA. Intensive,

continuing investigations in this basic area of research are strongly indicated.

In the past it has been assumed that the chemical carcinogens operate by removing the protective protein sheaths from the latent virus, releasing the infectious DNA in the cell. Although there is little direct evidence, it may well be that the overall transformation induced by the chemical agent in a less specific fashion and over a much longer time is closely related, in part of the process, to that accomplished more rapidly by the virus (or its bare DNA).

Energy Transfer and Carcinogenesis

Several investigators have attempted to associate the functioning of the carcinogenic polycyclic compounds with an energy transfer between the carcinogen and the DNA molecule at the genetic center of the somatic cell, leading to a chemical change which results in an altered and malignant DNA.[37] It has been shown by Szent-Gyorgyi and coworkers[39] that there is an apparent correlation between carcinogenic activity and the ability of some compounds to form charge-transfer complexes with molecular iodine.

Electron donation and acceptance by carcinogenic compounds has been studied by Allison and Nash[39a] for polycyclic hydrocarbons, steroids, polycyclic heterocyclics, and aromatic amines. Charge-transfer complex formation with chloranil in acetonitrile and with 1,3,5-trinitrobenzene in benzene, and electron spin-resonance in complexes with tetracyanoethylene were measured and the results found to show a broad general correlation (with noticeable discrepancies) between carcinogenicity and electron donation and acceptance.

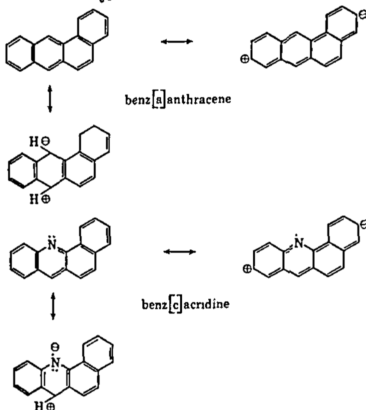
The simultaneous ability of many carcinogenic compounds to donate and accept electrons is ascribed to the quinoid or pseudoquinoid character of resonance for the polycyclic

compounds. Steric factors and the stability of the carcinogen under physiological conditions were also suggested as important factors.

It has been suggested[22, 39a, 40] that theory can be related to biological systems by means of a "sandwich"-type model wherein the near coplanar carcinogen (polycyclic compound) lies between two active sites in the metabolic system, one serving as an electron donor, the other as an electron acceptor, forming charge-transfer complexes with both. Allison and Nash suggest that this might[39a] be expected to cause the carcinogen to exert a primary effect on oxidative phosphorylation, leading to the altered DNA molecule. From a practical standpoint, their work[39a] clearly indicated that these tests demonstrating charge-transfer with trinitrobenzene and acridine would have found all of the major carcinogens studied. However, several noncarcinogens would also have been included in the positive group.

The Pullmans[39b] have attacked the theory of Allison and Nash[39a] and claim to have demonstrated that there is no correlation between the electron donating and accepting properties, and carcinogenicity. They cite examples of many compounds presenting ideal conditions for carcinogenicity according to the Allison and Nash hypothesis, which are devoid of carcinogenic activity. Allison and Nash[39c] have replied by pointing out that the energy coefficient measurements presented by the Pullmans are estimates rather than actual measurements of charge-transfer formation. Also, the original theory had taken into account the existence of several exceptions.

Sung[41] has attempted to relate absorption spectra and carcinogenicity. No proven carcinogen was found in 73 compounds with absorption spectra below 354.5 mμ, but 42 per cent of compounds with absorption spectra above 354.5 mμ (log $\epsilon > 1$) were carcinogenic, with λ_{max} between 354.5 and 368 mμ were carcinogens. Included in this latter



group were 3,4-benzpyrene, 3,4,8,9-dibenzpyrene, cholanthrene, 20-methylcholanthrene, various methylbenzanthracenes, and methylbenzacridines The results are thought to be in agreement with previous work by this author who demonstrated that, for 28 methylbenzacridines, 13/14 carcinogenic compounds have ΔE_2 below 1.132 kcal and 11/14 noncarcinogenic compounds have a ΔE_2 above this value.

Wentworth and Becker[42a] have investigated a method for determining the electron affinities of organic molecules which employs an electron capture cell used as a sensitive detector in gas chromatography They have discussed the

relationship between the electron absorption coefficient determined in this way and the electron affinity of the molecule. The electron affinities and ionization potentials of various aromatic hydrocarbons have been correlated.[42b] The stability of complexes formed between methylbenzenes and polycyclic aromatic hycrocarbons was found to be proportional to increased electron affinity to the polycyclic compounds, adding further proof to a charge-transfer complex interpretation.[42c] It would be interesting to attempt a correlation of electron affinities as determined with carcinogenic activity for a suitable series of compounds.

Huggins and Yang[40] studied the ability of a wide variety of carcinogenic and noncarcinogenic substances to induce mammary cancer in rats. They came to the conclusion that the molecular dimensions of the active polycyclic hydrocarbons had a close resemblance to steroids insofar as the planar dimension was concerned, but their thickness was less, being 3.6Å as compared with 5.6Å for the noncarcinogenic, nonplanar steroidal compounds. It was suggested that these thin, flat molecules containing four to five condensed rings could be expected to interact with base-pairs (guanine-cytosine or adenine-thymine) in the Watson-Crick model of DNA, leading to a modification and carcinogenesis.

Boyland and Green[21] and Liquori and DeLerma[22] have recently reported *in vitro* reactions between DNA and benzpyrene, dibenzanthracene, and pyrene. It is suggested by both groups that the hydrocarbons intercalate between base pairs in the DNA molecule. Results have been reported by the latter group,[22] of solubilization experiments of benzpyrene and dibenzanthracene in DNA-water solutions in conjunction with observed changes in ultraviolet absorption spectra and fluorescence spectra of the dissolved benzpyrene. It is suggested that the aromatic molecules become inserted between base-pairs in the "disordered," stretched sections of the DNA molecule. The nature of the sites is not specified,

but it is strongly suggested that the empty spaces between occasional bases, present in very small numbers, may exist in "undenatured" DNA and may be filled by the aromatic hydrocarbons. These may increase on denaturation of DNA, on dilution, or on heating. Weak charge-transfer and dipole-induced dipole–dipole forces between the polar purine bases and the polarizable aromatic molecules may stabilize the complexes. Lerman[24a] used studies of sedimentation, low-angle X-ray scattering, flow dichroism, flowpolarized fluorescence, and chemical reactivity to obtain evidence for intercalation of acridines between two otherwise sequential base-pairs in DNA. The binding is thought to require a local untwisting and extension of the double helix. Although intercalation seems to be a prerequisite for mutagenicity of the insertion–deletion type, the acridine structure is not essential nor are all intercalating molecules found to be mutagenic.

Brookes and Lawley[24b] studied the binding of a variety of tritium-labeled hydrocarbons with RNA, DNA, and protein in mouse skin. In one case, a C^{14}-labeled compound was used. No binding occurs immediately after application, and maximal binding takes place after 24–48 hours. The results of a large number of experiments show that the partition of each individual hydrocarbon among the cellular constituents, RNA, DNA, and protein gave a consistent pattern which differed for the various hydrocarbons in a fashion related to their carcinogenic potency as expressed by the Iball index.[43] The extent of binding to total protein or RNA showed no correlation in this regard, but binding to DNA showed a significant positive correlation with carcinogenic potency. It is suggested by Brookes and Lawley[24b] that DNA is the essential cellular receptor of the carcinogens. The nature of the binding is not clear, but it is suggested that a metabolite of the hydrocarbon (i.e. an epoxide) is actually the bound form.

It seems probable that many compounds can be expected to form charge-transfer complexes with DNA which are not carcinogens. The important factor may be whether or not the charge-transfer complex then proceeds to form an addition product with DNA in the case of carcinogens.

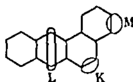
Heidelberger[24c] admits that DNA can be shown to bind to aromatic hydrocarbons but claims there is no correlation with carcinogenicity, since no localization of radioactivity in the nuclei could be detected in C^{14}-containing hydrocarbons. He has shown that the Boyland[21] and Liquori[22] experiments are probably invalid. Actually, a colloidal suspension between the hydrocarbon particles and the DNA seems to form rather than an intercalated molecular complex. The hydrocarbon seems to stabilize the DNA colloid to some extent. Also the Boyland DNA was mainly single strand rather than double strand.

The K-L-M *Region Theory of Carcinogenesis*

Although the precise details of the interactions of tissue with carcinogens at the molecular level are still not known with certainty, considerable progress in this direction has been made since the important pioneering work by the French group of workers, reviewed by the Pullmans in 1955.[41] This school developed an electronic theory based upon molecular orbital calculations of activation energies to predict which polycyclic hydrocarbon should be the most carcinogenic. They concluded that, in carcinogenic molecules, the energy of the phenanthrenoid bond, designated as the K (Krebs) region must not exceed a certain limit, while the activation energy of the anthrenoid or meso (L) region of the molecule must exceed another value.

It now appears that the statement of the original theory was an oversimplification of the complex structural require-

ments for carcinogenesis. Some recent elaborations of this theory, taken in conjunction with a consideration of steric factors and the physical properties of the molecules, are more encouraging and suggestive of further fundamental studies.



On the basis of new structural variation studies and on previous data, the role of the K-region of polycyclic compounds in carcinogenesis has been recently reviewed by the Millers[45] and the theory made more definitive. For example while substitution of fluorine in the 5-position of 7-methylbenz[a]anthracene destroys carcinogenicity, the 6-fluoro derivative has nearly the same activity as the parent compound. These and other data suggest that, of the two K-region positions, availability of an unsubstituted 5-position is far more important than a free 6-position.

In summing up their studies with the angular benzacridines, Lacassagne and co-workers[17] concluded in 1956: "it now seems certain that the greater or lesser carcinogenicity of a molecule is linked to very slight structural differences, and it is probable that with only very refined techniques will there be success in understanding these differences. The important steric effects—which govern the affinity of the molecule to its biological carrier—must not be overlooked. These effects are clearly demonstrated by the disappearance of carcinogenicity when side chains on the hydrocarbon are lengthened (a phenomenon which has been pointed out in the angular benzacridines)" (see p. 367 of Ref. 17). Thus, eight years ago, the French workers recognized the importance of the major parameters involved in carcinogenic

activity of polycyclic compounds, which have recently been restated and elaborated by new workers entering this field. With the Hammett substituent constant, σ, and a substituent constant, π, defined as $\pi = \log Px - \log P_H$ (P_H is the partition coefficient of a parent compound and Px that of a derivative), regression analyses have been made of the effect of substituents on the biological activity of carcinogenic compounds on mice. The results clearly indicate that the lipohydrophilic character of these molecules should be considered in attempts to rationalize structure–activity relationships. Specifically, these studies have led to a simple explanation for the inactivity of the large molecules of six or more rings and the higher alkyl derivatives of 1,2-benzanthracene. The large inactive molecules are generally too lipophilic (larger values of log P), while small inactive molecules are not lipophilic enough (low values of log P) (Hansch and Fuita[46]). They suggest that this parameter is a more fundamental one to consider than variations in steric factors or the importance of K- and L-regions in these molecules.

Chemical Reactivity and Carcinogenicity of Polycyclic Compounds

In general, both polycyclic aromatic hydrocarbons and heterocyclics show a low order of chemical reactivity under laboratory conditions of pH and temperature approximating physiological conditions. Thus, except for charge-transfer complex formation, discussed previously, attempts to relate carcinogenicity of a series to the various members' ability to react with various standard reagents have met with only limited success.

In 1954, Badger[47] suggested that three reagents which add to bonds (i.e. the K-region) rather than by replacing functional groups might prove useful. These reagents were ozone, diazoacetic acid, and osmium tetroxide. Badger

reported a "very satisfactory correlation" between the rate of addition of osmium tetroxide and the bond order at the K-region, calculated by the molecular orbital method for a group of nonsubstituted aromatic polycyclic compounds. The agreement was very good when the rates of reaction were compared with the Pullmans' electronic index. Thus, there is good agreement between the excess charge at the K-region and the rate of reaction with osmium tetroxide, and it is thus shown that the osmium tetroxide reaction confirms the relationship between the calculated electronic indices for the K-region and the rate of reaction with a reagent which attacks this part of the molecule

Badger found that both electron-repelling (methyl) and electron-attracting (cyano) groups convert benz(a)anthracene into carcinogenic material, one might note that some of the electron-attracting groups may engage in H-bonding with centers in DNA The OsO_4 results indicated that the more carcinogenic materials often reacted the more rapidly with OsO_4, which in turn reacts with those aromatic hydrocarbons having the greater charge at the K-region as calculated by either the valence bond or molecular orbital method

The ozonization of polycyclic aromatic hydrocarbons has been extensively investigated by Moriconi,[48] who has shown that ozone can react at all three of the reaction sites designated by the Pullmans as the K-, L-, and M-regions Attempts have been made to correlate K-, L-, and M-region activity of both carcinogenic and noncarcinogenic compounds with the path and mechanism of the ozonolysis reaction It turns out that ozone attacks the potent carcinogens mainly at the L-region rather than at the K-region

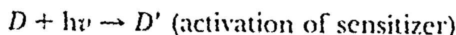
It seems probable that the metabolic reactions likely to be related to the course of these ozone reactions are the reactions of detoxification rather than the charge-transfer complex formation with tissue constituents most likely to be responsible for the carcinogenicity in these series

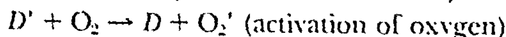
nonnecrotizing dose of DMBA protects the animal from the necrosis-inducing effect of larger amounts of the same compound. These studies indicated that the presence of a critical amount of adrenocortical steroids in the adrenal gland was a prerequisite to the induction of necrosis by DMBA.

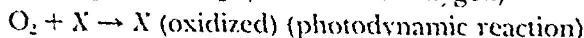
Photodynamic Activity and Carcinogenic Action

As long ago as 1900, Raab[54] found that the time required for various dyes to kill paramecia depended on the light intensity. Since then, dyes and related substances have been referred to as "sensitizers" in the light-catalyzed processes designated as "photodynamic." This field was summarized up to 1941 by Blum.[55] The chemical reaction which occurs in these processes appears to be a photosensitized oxidation by the dissolved oxygen. Doniach and co-workers[56] reported that carcinogenic hydrocarbons were 10^3 to 10^4 times as active for killing paramecia as the dyes which had normally been used as photodynamic sensitizers.

Koffler and Markert[57] have shown that DNA is degraded photodynamically by methylcholanthrene and 1,2-benzanthracene. It is thus apparent that photodynamic action can degrade macromolecules as does ionizing radiation, and the carcinogenic hydrocarbons are unusually effective sensitizers for these reactions. Blum[55] has described the process by the following steps:


$D + h\nu \rightarrow D'$ (activation of sensitizer)


$D' + O_2 \rightarrow D + O_2'$ (activation of oxygen)


$O_2 + X \rightarrow X$ (oxidized) (photodynamic reaction)

Where D is the sensitizer and X the substrate (e.g. macromolecule reacting.

Mottram and Doniach[56a] reported a rather impressive correlation between photodynamic toxicity and carcinogenicity.

The Doniach test system has been used by Epstein and co-workers[58] to study photodynamic toxicity in several series of polycyclic aromatic hydrocarbons and polycyclic heterocyclic compounds with wide structural variations. A general correlation between photodynamic toxicity and carcinogenicity has been found. The test system was also successfully applied to the bioassay of pure samples of 3,4-benzpyrene and some other polycyclic carcinogens in the concentration range of $10^{-5} - 10^{-10}$ gm/ml.[58b]

The photodynamic response of paramecium to 3,4-benzpyrene appears to be oxygen-dependent, and marked protection is afforded by antioxidants such as oxotocopherol and butylated hydroxy anisole when added to the incubation system.[58d] The sensitivity of the response could also be modified by a variety of other factors such as tryptophan, fractions of plasma proteins, nonionic wetting agents, and inorganic salts. No evidence for the involvement of $-SH$ groups or of peroxide formation could be demonstrated in the photodynamic process 3,4-Benzpyrene, or its presumed water-soluble photodynamically active product, was found not to be firmly bound to *Paramecium caudatum* [58d]

After studying the photodynamic response of 157 carcinogenic and noncarcinogenic polycyclic aromatic hydrocarbons and heterocyclic compounds, Epstein and co-workers[58e] came to the following general conclusions in connection with the relationships among structure, carcinogenicity, photodynamic activity, and light absorptivity.

1 Significant absorption of light from the experimental irradiation system was shown to be a prerequisite but not sufficient requirement for high photodynamic activity
2 High photodynamic activity is largely confined to 4- and 5-ring compounds whether homocyclic or heterocyclic
3 Inactivity cannot be ascribed to difficulties in compound solubilization or cell uptake alone

4. A statistically significant association between photodynamic activity and carcinogenicity was demonstrated.
5. With a defined threshold, it was shown that compounds with high photodynamic activity are likely to be three times as carcinogenic as compounds with low activity, but the photodynamic assay cannot identify a particular compound as being carcinogenic or noncarcinogenic.

It appears that the preliminary use of the photodynamic assay method would select a *passing group* for which the odds of carcinogenic activity would be three times as great as for compounds in the *failing group.*

Concerning the effect of structural variations, high photodynamic activity, like strong carcinogenicity, seems to be limited to compounds with four or five rings. Charge-transfer complex formation may be a common factor to both carcinogenicity[37–40] and photodynamic response. Charge-transfer complex formation with tissue constituents such as DNA[22] or proteins[18] could account for the observation of the critical molecular geometrical requirements for the polycyclic compounds both for carcinogenicity[40] and for photodynamic toxicity.[58e] However, Epstein and co-workers have now found that there is no direct correlation between carcinogenicity of a series of aromatic compounds and their involvement in forming charge-transfer complexes with iodine, chloranil, trinitrobenzene, or acridine.[58f]

Photodynamic assay has been applied to the assay of crude organic extracts such as atmospheric pollutants.[58b] This application is based[58e] upon the fact that the principal known carcinogen in the atmosphere is benzo(a)pyrene, which is strongly photodynamically active, and, generally in industrial areas, there is a high correlation between the amount of this strong carcinogen and other polycyclic hydrocarbons in the air.

In general it was found that amino groups on the 1-position

of pyrene, 6-position of chrysene (6-aminochrysene inhibits growth of induced tumors,[51] and 4-position of fluoranthene markedly increased photodynamic response of the inactive or slightly active parent compounds. Both methyl and especially amino groups on the 6-position of benz(a)pyrene result in a marked decrease in photodynamic activity as when these same groups are introduced into the 7-position of benz(a)-anthracene

Among a series of benz(c)acridines prepared by Cromwell and co-workers[15] and tested by Epstein,[58e] there is a marked increase in activity of the 7,9-dimethyl substituent with successive decreases for the 5,6-dimethyl,7,10-dimethyl, and 1,4-dimethyl derivatives

The 5,6-dimethyl structure shows increased activity with chlorine in the 7-position, decreased with chlorine in the 10-position, and decreased further with 9- or 11-chloro substituents It is interesting that, for the dimethylbenz(c)-acridine series, the carcinogenicity is known to be strong for the 7,9-derivative, generally weak for the 5,6-derivatives, and weakest for the 5,6-dimethyl-11-chlorobenz(c)-acridine [59] The carcinogenicity of the new [60] 7-chloro derivative has not yet been measured

In contrast with the high photodynamic activity for the 7-chloro-5,6-dimethylbenz(c)acridine, there is a successive decrease in activity for the corresponding 7-morpholino, 7-phenoxy, and 7-acetoxy derivatives [60] This reduction in activity may be assigned to the increased steric requirements for these larger groups It was also found that the 5,6-dimethylbenz(c)acridone[58] shows no photodynamic activity

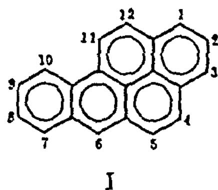
Another series for which carcinogenicity studies are as yet unavailable are the 11H-indeno-quinoxalines [16c] The parent compound shows strong photodynamic activity with decreasing activity for the 11-hydroxy-11-methyl and 5,10-di-N-oxo-11-methyl-11-hydroxy derivatives respectively The 11-one and 5-N-oxo-11-one derivatives have a similar degree

of photodynamic activity to the 11-hydroxy-11-methyl derivative. On the basis of these studies, it is seen that a K-region does not appear to be necessary for photodynamic activity, and linear as well as angular compounds are found to be active. It seems of interest to study the carcinogenicity of some of the indenoquinoxalines.

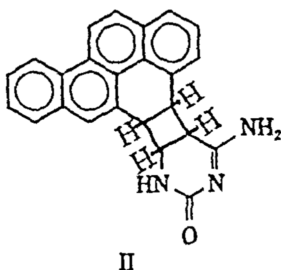
From the *in vivo* data reported by Epstein,[58] it appears that significant light absorption is a prerequisite but not sufficient for photodynamic activity. The exact location of absorption peaks does not appear to be important, and fluorescence is inconsequential.

The work of Rice[23] shows clearly that benzo(a)pyrene forms stable addition products with uracil, thymine, cytosine, 5-methylcytosine, guanine, and 6-azothymine upon irradiation with ultraviolet light above 320 mμ in aqueous solution at pH 6–7.

It is suggested that the general structure of the benzpyrene-pyrimidine adducts are as indicated in structure II.



I



II

These workers hope to be able to correlate carcinogenic activity of various hydrocarbons with their ability to form photo-adducts in 4 per cent aqueous sodium dodecyl sulfate solutions. It would seem to be of more importance to seek a correlation between carcinogenicity and the dark reaction of a series of carcinogens with these biologically important bases.

Syntheses of New Polycyclic Heterocyclic Compounds of Biological Interest

A working hypothesis for carcinogenesis and carcinostasis to guide further research with the polycyclic compounds is based on the following multifacet rationale. The carcinogen initiates an infectious DNA through a dual interaction with the cell nucleus. The nucleoprotein complexes with the polycyclic compound, and the DNA is laid bare for further attack. Intercalation of the large flat polycycles between base-pairs in the more disordered stretched sections of the DNA chain follows. Such intercalation involves charge-transfer complexation (electron donation and acceptance) which would be expected to provide increased stability to such disordered sections of the DNA through π-orbital overlap. Such specific complex formation might mask part of the nucleic acid chain (some of the bases) in the case of compounds with critical steric requirements, leading to interference with the replication process for the normal DNA and alteration in the assembling of amino acids in the nucleoprotein by messenger RNA. Thus, a new cell nucleus emerges. When it happens that there is no controlling immune response available from the host, uncontrolled growth begins and neoplasia is started which no longer requires the presence of the original carcinogen for its propagation. It is also possible that the carcinogen plays an important role in the suppression of the normal immune response expected from the host to deal with appearance of strange cell nuclei. Complex formation between the carcinogen and specific antigens might result in their loss by the animal at this critical time, allowing for the uncontrolled growth of the abnormal cell.

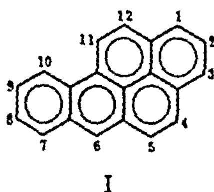
Concentration factors are of importance in the related process of carcinogenesis versus carcinostasis, see Refs. 30 and 51. Thus, it is suggested that chemically induced

of photodynamic activity to the 11-hydroxy-11-methyl derivative. On the basis of these studies, it is seen that a K-region does not appear to be necessary for photodynamic activity, and linear as well as angular compounds are found to be active. It seems of interest to study the carcinogenicity of some of the indenoquinoxalines.

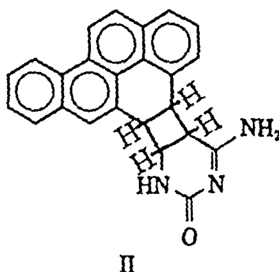
From the *in vivo* data reported by Epstein,[58] it appears that significant light absorption is a prerequisite but not sufficient for photodynamic activity. The exact location of absorption peaks does not appear to be important, and fluorescence is inconsequential.

The work of Rice[23] shows clearly that benzo(a)pyrene forms stable addition products with uracil, thymine, cytosine, 5-methylcytosine, guanine, and 6-azothymine upon irradiation with ultraviolet light above 320 mμ in aqueous solution at pH 6–7.

It is suggested that the general structure of the benzpyrene-pyrimidine adducts are as indicated in structure II.



I



II

These workers hope to be able to correlate carcinogenic activity of various hydrocarbons with their ability to form photo-adducts in 4 per cent aqueous sodium dodecyl sulfate solutions. It would seem to be of more importance to seek a correlation between carcinogenicity and the dark reaction of a series of carcinogens with these biologically important bases.

Syntheses of New Polycyclic Heterocyclic Compounds of Biological Interest

A working hypothesis for carcinogenesis and carcinostasis to guide further research with the polycyclic compounds is based on the following multifacet rationale. The carcinogen initiates an infectious DNA through a dual interaction with the cell nucleus. The nucleoprotein complexes with the polycyclic compound, and the DNA is laid bare for further attack. Intercalation of the large flat polycycles between base-pairs in the more disordered stretched sections of the DNA chain follows. Such intercalation involves charge-transfer complexation (electron donation and acceptance) which would be expected to provide increased stability to such disordered sections of the DNA through π-orbital overlap. Such specific complex formation might mask part of the nucleic acid chain (some of the bases) in the case of compounds with critical steric requirements, leading to interference with the replication process for the normal DNA and alteration in the assembling of amino acids in the nucleoprotein by messenger RNA. Thus, a new cell nucleus emerges When it happens that there is no controlling immune response available from the host, uncontrolled growth begins and neoplasia is started which no longer requires the presence of the original carcinogen for its propagation It is also possible that the carcinogen plays an important role in the suppression of the normal immune response expected from the host to deal with appearance of strange cell nuclei Complex formation between the carcinogen and specific antigens might result in their loss by the animal at this critical time, allowing for the uncontrolled growth of the abnormal cell

Concentration factors are of importance in the related process of carcinogenesis versus carcinostasis; see Refs 30 and 51. Thus, it is suggested that chemically induced

immunization against certain forms of cancer may become a reality.

During the course of a general program involving the synthesis of potential carcinogenic (cancer-forming) and carcinolytic (antitumor) agents, new methods of synthesis were developed which provide for a wide variety of new types of derivatives of the biologically important benzacridines[15] and the less well-known indenoquinolines,[16a, b] indenoquinoxalines,[16c] and the benzophenazines.[16d] The biological interest in compounds of this nature may be said to have developed when it was discovered in France[17] that benzacridines of the angular type caused cancer on the skin of inbred strains of mice when properly applied.

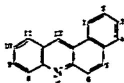
These new compounds are intended for testing both as carcinogenic and antitumor agents, although, as a class, they have been mainly recognized in the past as carcinogenic substances. Since most known antitumor agents have also been found to be carcinogenic in action with the proper system, it seemed logical to us that series which have essentially been known for their carcinogenic action might include some examples that would prove to be effective antitumor agents. Thus, by subtle changes in chemical structure, one might be able to convert a series of compounds from essentially carcinogenic substances to mainly antitumor agents.

The relationship between chemical constitution and carcinogenic activity (and the related process of carcinostasis) for these several series of large flat molecules is becoming clearer, and it is hoped that further work will provide for the elucidation of these important biological processes.

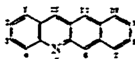
It seems obvious that, if the organic chemist is to help solve these related problems of carcinogenesis and carcinolytic action of various agents, he must play a role far broader than merely supplying new agents showing these effects. In the years ahead, he must bring to bear upon these related problems his understanding of the mechanisms

Polycyclic Heterocyclic Compounds

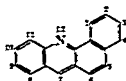
The Benzacridines



(a)

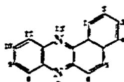


(b)

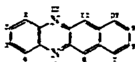


(c)

The Benzophenazines

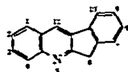


(a)

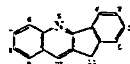


(b)

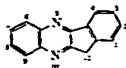
The Indenoquinolines and Indenoquinoxalines



6H—(2,1-b)—



11H—(1,2-b)—



11H—(1,2-b)—

17. A. Lacassagne, N. P. Hoi-Buu, R. Daudel, and F. Zazdela, The Relation Between Carcinogenic Activity and the Physical and Chemical Properties of Angular Benzacridines, *Advan. Cancer Res. 4*, 315 (1956).

18. *See:* The Relation of Protein Binding to Hydrocarbon Carcinogenesis, by C. Heidelberger, p. 179-92, in Ref. 12; and C. W. Abell and C. Heidelberger, The Interaction of Carcinogenic Hydrocarbons with Tissues. VIII. Binding of Tritium-labeled Hydrocarbons to the Soluble Proteins of Mouse Skin, *Cancer Res., 22*, 931 (1962).

19. See Chap. 15, Ref. 11.

20. P. Alexander and K. A. Stacey, Comparison of the Changes Produced by Ionizing Radiations and by Alkylating Agents; Evidence for a Similar Mechanism at the Molecular Level, *Ann. N.Y. Acad. Sci., 68*, 1225 (1958); see also p. 195 of Ref. 12.

21. E. Boyland and B. Green, The Interaction of Polycyclic Hydrocarbons and Nucleic Acids, *Brit. J. Cancer, 16*, 507 (1962).

22. A. M. Liquori et al., Interaction Between DNA and Polycyclic Aromatic Hydrocarbons, *J. Mol. Biol., 5*, 521 (1962).

23. J. M. Rice, Photochemical Addition of Benzo(a)pyrene to Pyrimidine Derivatives, *J. Am. Chem. Soc., 86*, 1444 (1964).

24. (*a*) L. S. Lerman, Acridine Mutagens and DNA Structure, *J. Cell. Comp. Physiol., 64* Suppl. 1, 1 (1964).
  (*b*) P. Brookes and P. D. Lawley, Reaction of Some Mutagenic and Carcinogenic Compounds with Nucleic Acids, ibid., p. 111.
  (*c*) C. Heidelberger, Studies on the Molecular Mechanism of Hydrocarbon Carcinogenesis, ibid., p. 129.

25. (*a*) E. C. Miller and J. A. Miller, In Vivo Combinations between Carcinogens and Tissue Constituents and Their Possible Role in Carcinogenesis, *Cancer Res., 12*, 547 (1952).
  (*b*) C. Heidelberger, Immunological Problems and Techniques, *Cancer Chem. Reports, 20*, 19 (1962).

26. See Ref. 11, Chap. 17.

27. H. N. Green, An Immunological Concept of Cancer: A Preliminary Report, *Brit. Med. J., ii*, 1374 (1954).

28. E. A. Kobat, Size and Heterogenicity of the Combining Sites on an Antibody Molecule, *J. Cell. Comp. Physiol., 50*, Suppl. 1, 79 (1957).

29. Ref. 11, p. 432.

30. T. L. DAO and Y. TANAKA, Inhibitory Effect of Polynuclear Hydrocarbons and Amphenone Analogs on Induction of Acute Adrenal Necrosis by 7,12-Dimethylbenz(a)anthracene, *Cancer Res.*, *23* (8) 1148 (1963)

31. (*a*) E. BOYLAND, The Mechanism of Tumor Induction by Aromatic Amines and Other Carcinogens, *Z. Krebsforsch.*, *65* (4), 378 (1963)

(*b*) E. BOYLAND, C. E. DUKES, and P. L. GROVER, Carcinogenicity of 2-Naphthylhydroxyl Amines, *Brit. J. Cancer*, *17* (1), 78 (1963)

32. (*a*) J. W. CRAMER, J. A. MILLER, and E. C. MILLER, N-Hydroxylation: A New Metabolic Reaction Observed in the Rat with the Carcinogen 2-Acetylaminofluorene, *J. Biol. Chem.*, *235*, 885 (1960)

(*b*) E. C. MILLER, J. A. MILLER, AND H. A. HARTMAN, N-Hydroxy-2-acetylaminofluorene: A Metabolite of 2-Acetylamino-fluorene with Increased Carcinogenic Activity in the Rat, *Cancer Res.*, *21*, 815 (1961)

33. D. B. CLAYSON and M. J. ASHTON, The Metabolism of 1-Naphthylamine and Its Bearing on the Mode of Carcinogenesis of the Aromatic Amines, *Acta IXn Int. Cancer Congress*, *19* (3-4), 539 (1963)

34. P. H. WEISBURGER, P. H. GRANTHAM, and J. H. WEISBURGER, Differences in the Metabolism of N-Hydroxyl-N-2-fluorenylacetamide in Male and Female Rats, *Biochemistry*, *3*, 808 (1964)

35. P. KOTIN and H. L. FALK, Organic Peroxides, Hydrogen Peroxide, Epoxides and Neoplasia, *Radiat. Res.*, *Suppl. 3*, 193 (1963)

36. *Progress Report XI*, *Virus and Cancer*, pp. 54-56, Sloan-Kettering Institute for Cancer Research, New York (Jan. 1963)

37. N. ARLEY and R. EKER, Mechanisms of Carcinogenesis, *Advan. Biol. Med. Phys.*, *8*, 375 (1962)

38. A. SZENT-GYORGYI, I. ISENBERG, and S. L. BAIRD, JR., On the Electron Donating Properties of Carcinogens, *Proc. Nat. Acad. Sci. U.S.*, *46*, 1444 (1960)

39. (*a*) A. C. ALLISON and T. NASH, Electron Donation and Acceptance by Carcinogenic Compounds, *Nature*, *197*, 758 (1963)

(*b*) B. PULLMAN and A. PULLMAN, Electronic Aspects of the Interactions Between the Carcinogens and Possible Cellular Sites of Their Activity, *J. Cell. Comp. Physiol.*, *64 Suppl. 1*, 91 (1964)

(*c*) A. C. ALLISON and T. NASH, *Nature*, *199*, 471 (1963)

40. C. HUGGINS and N. C. YANG, Induction and Extinction of Mammary Cancer, *Science*, *137*, 257 (1962).

41. SHOU-SIN SUNG, Relationship Between Carcinogenic Power and Electronic Transition Energy, *C. R. Acad. Sci. (Paris)*, *257* (6), 1425 (1963).

42. (*a*) W. E. WENTWORTH and R. S. BECKER, Potential Method for the Determination of Electron Affinities of Molecules: Application to Some Aromatic Hydrocarbons, *J. Am. Chem. Soc.*, *84*, 4263 (1962).
(*b*) R. S. BECKER and W. E. WENTWORTH, Electron Affinities and Ionization Potentials of Aromatic Hydrocarbons, ibid., *85*, 2210 (1963).
(*c*) W. E. WENTWORTH and E. CHEN, Molecular Interaction Between Methylbenzenes and Polycyclic Aromatic Hydrocarbons, *J. Phys. Chem.*, *67*, 2201 (1963).

43. J. IBALL, The Relative Potency of Carcinogenic Compounds, *Am. J. Cancer*, *35*, 188 (1939).

44. A. PULLMAN and B. PULLMAN, Electronic Structure and Carcinogenic Activity of Aromatic Molecules, *Advan. Cancer Res.*, *3*, 117 (1965).

45. J. A. MILLER, and E. C. MILLER, The Carcinogenicities of Fluoro Derivatives of 10-Methyl-1,2-benzanthracene. II. Substitution of the K-Region and the 3′-, 6-, and 7-Positions, *Cancer Res.*, *23* (2) (Pt. 1), 229 (1963).

46. C. HANSCH, and T. FUITA, φ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structures, *J. Am. Chem. Soc.*, *86*, 1616 (1964).

47. G. M. BADGER, Chemical Constitution and Carcinogenic Activity, *Advan. Cancer Res.*, 2, 73 (1954).

48. E. J MORICONI, and L. B. TARANKO, Ozonolysis of Polycyclic Aromatics. XI. 3-Methylchlolanthrene, *J. Org. Chem.*, *28*, 2526 (1963), and preceding papers in the series.

49. E. L. WYNDER and D. HOFFMANN, The Carcinogenicity of Benzofluoranthenes, *Cancer* (Phila.), *12*, 1194 (1959).

50. H. N. GREEN, Immunological Aspects of Cancer, in *Cancer*, Vol. 3, 1, ed. Raven, Butterworth, London (1958).

51. N. P. BUU-HOI, Polycyclic Aromatic Hydrocarbons, Their Analogues and Derivatives as Potential Pharmacological Agents, *Med. Exp.* (Basel), *8* (4-6), 209 (1963).

52. See Fig. 1, p. 227, Ref. 14.

53. See Table IV, p. 138, Ref. 14.

54. O. RAAB, *Z. Biol.*, *39*, 524 (1900).

55 H. F. Blum, *Photodynamic Action and Diseases Caused by Light*, Reinhold, New York (1941).

56 (*a*) J. C. Mottram and I. Doniach, *Nature*, *140*, 588 (1937); *Lancet*, *234*, 1156 (1938)
(*b*) I. Doniach, *Brit. J. Exp. Path.*, *20*, 227 (1939).

57 H. Koffler and I. L. Markert, *Proc. Soc. Exp. Biol. Med.*, *76*, 90 (1951)

58 (*a*) S. S. Epstein and M. Burroughs, Some Factors Influencing the Photodynamic Response of *Paramecium caudatum* to 3,4-Benzpyrene, *Nature*, *193*, 337 (1962)
(*b*) S. S. Epstein et al., Photodynamic Bioassay of Benzo(a)pyrene with *Paramecium caudatum*, *J. Nat. Cancer Inst.*, *31*, 163 (1963)
(*c*) S. S. Epstein, The Photodynamic Activity of Polycyclic Hydrocarbon Carcinogens, *Ext. de Acta U.I.C. le C. XIX*, (3-4), 599 (1963)
(*d*) S. S. Epstein, M. Burroughs, and M. Small, The Photodynamic Effect of the Carcinogen 3,4-Benzpyrene, on *Paramecium caudatum*, *Cancer Res.*, *23*, (1), 35 (1963)
(*e*) S. Epstein, M. Small, H. L. Falk, and N. Mantel, The Correlation Between Photodynamic and Carcinogenic Activities in Polycyclic Compounds, ibid., *24*, (5), 855 (1964)
(*f*) S. S. Epstein, I. Bulon, J. Koplan, M. Small, and N. Mantel, Charge-Transfer Complex Formation, Carcinogenicity, and Photodynamic Activity of Polycyclic Compounds, *Nature*, *204*, 750 (1964)

59 Private communication to N. H. Cromwell by R. K. Boutwell, McArdle Memorial Laboratory, University of Wisconsin, Madison, Aug. (1960)

60 N. H. Cromwell and I. A. Nielsen, unpublished

5. COLOR NAMES FOR COLOR SPACE

By Alphonse Chapanis
The Johns Hopkins University

The richness, variety, and importance of normal color experience are truly remarkable. Our ability to see colors contributes unmeasurably to our ideas of beauty and to the aesthetic appreciation of objects in our everyday world. Color is used routinely in business, industry, science, and medicine to code and identify objects and to communicate information. And, finally, color seems to be associated with a variety of affective responses, feelings, emotions, and moods—such as liking, disliking, excitement, and depression.

For these reasons, it is not surprising to find that there has been a great deal of scientific work done on color vision and color perception. Richter's two bibliographies,[1,2] for example, contain well over 6,000 entries even though they cover only the fifteen-year period from 1940 through 1954. Despite all this work, there are many fundamental things we still do not know about this subject. As one illustration, there is still much uncertainty about precisely how various wavelengths of radiant energy are transformed into nervous energy and differentially coded for transmission to the higher nerve centers of the brain.

Color can be studied anywhere along a broad spectrum of problems. At one extreme, some scientists are deeply interested in the microfunctioning of the retinal elements and in

the photochemicals that transform electromagnetic energy into nervous energy. At the other extreme, one can find research workers concerned with responses to color as an aid to understanding the structure and dynamics of personality. The problem I wish to discuss falls somewhere near the middle of this spectrum, for I am primarily concerned with the common color names that we can all use to talk about our color sensations.

Strangely enough, in all the vast literature on color and color perception, the topic of color names is one that psychologists have largely ignored. To be sure, studies involving color naming have had a long and respectable history in psychology. In fact, one of the very earliest of these was begun by James McKeen Cattell while he was a fellow at The Johns Hopkins University in 1882-83.[3] But Cattell and the many psychologists who used color-naming tests after him were interested in these tests almost exclusively as dependent variables. Typically, investigators have used the time taken to name a set of simple colors, or the number of errors made in naming some simple colors, as a measure of other things—mental ability, the effect of drugs, the effect of fatigue and so on (see, for example, Berdie,[4] Cattell,[5] Cattell and Farrand,[6] Remoli[7]). Rarely was anyone interested in what people meant by the color names they were using. The occasional studies that tried to establish some denotative meaning for color names confined themselves to the spectrum colors (see, for example, Beare,[8] Dimmick and Hubbard,[9] and Pickford[10]), to very few surface colors (Katzin and Murray[11]), or to no more than five or six simple color names (see, for example, Halsey[12]).

To focus more precisely on the present topic, let me pose the question that got me started on this problem in the first place: What is the maximum number of usable color names for all of color space?

Why Study Color Names?

Let's begin by asking, "Why should anyone study color names?" Although there are several answers to this question, one of the simplest is to say that colors and color names are important for many industrial and engineering purposes. You have heard, I know, that we are experiencing a kind of second industrial revolution—a revolution brought about by new kinds of machines and machine systems. One thing that characterizes this new class of machines is that they deal not so much with tangible products, like steel ingots, glass bottles, and washing machines, as with information—meaningful information that can be understood, handled, and transferred from man to machine and vice versa

Let me give you an example An air traffic control center in a large metropolitan airport doesn't produce a single concrete item that one can point to, handle, or put into a crate Its principal product is information—information about incoming flights, outgoing flights, weather conditions, and emergencies For any one aircraft the controller needs to know a lot of different things the identity of the aircraft, its altitude, speed, distance, direction, origin, and destination. Some of this information changes continually; other pieces of the information are relatively static Since controllers in a large metropolitan airport may have to keep tabs on a hundred or more aircraft at any one time, you can readily see that such a system must be capable of handling a very considerable amount of information

Another familiar example of modern systems that deal primarily with information are the giant computers that are becoming almost as common as desk calculators and slide rules were a few decades ago

An essential part of many of these information-handling systems is that they must portray, or display, the information

the photochemicals that transform electromagnetic energy into nervous energy. At the other extreme, one can find research workers concerned with responses to color as an aid to understanding the structure and dynamics of personality. The problem I wish to discuss falls somewhere near the middle of this spectrum, for I am primarily concerned with the common color names that we can all use to talk about our color sensations.

Strangely enough, in all the vast literature on color and color perception, the topic of color names is one that psychologists have largely ignored. To be sure, studies involving color naming have had a long and respectable history in psychology. In fact, one of the very earliest of these was begun by James McKeen Cattell while he was a fellow at The Johns Hopkins University in 1882-83.[3] But Cattell and the many psychologists who used color-naming tests after him were interested in these tests almost exclusively as dependent variables. Typically, investigators have used the time taken to name a set of simple colors, or the number of errors made in naming some simple colors, as a measure of other things—mental ability, the effect of drugs, the effect of fatigue and so on (see, for example, Berdie,[4] Cattell,[5] Cattell and Farrand,[6] Remoli[7]). Rarely was anyone interested in what people meant by the color names they were using. The occasional studies that tried to establish some denotative meaning for color names confined themselves to the spectrum colors (see, for example, Beare,[8] Dimmick and Hubbard,[9] and Pickford[10]), to very few surface colors (Katzin and Murray[11]), or to no more than five or six simple color names (see, for example, Halsey[12]).

To focus more precisely on the present topic, let me pose the question that got me started on this problem in the first place: What is the maximum number of usable color names for all of color space?

Why Study Color Names?

Let's begin by asking, "Why should anyone study color names?" Although there are several answers to this question, one of the simplest is to say that colors and color names are important for many industrial and engineering purposes. You have heard, I know, that we are experiencing a kind of second industrial revolution—a revolution brought about by new kinds of machines and machine systems. One thing that characterizes this new class of machines is that they deal not so much with tangible products, like steel ingots, glass bottles, and washing machines, as with information—meaningful information that can be understood, handled, and transferred from man to machine and vice versa

Let me give you an example. An air traffic control center in a large metropolitan airport doesn't produce a single concrete item that one can point to, handle, or put into a crate. Its principal product is information—information about incoming flights, outgoing flights, weather conditions, and emergencies. For any one aircraft the controller needs to know a lot of different things. the identity of the aircraft, its altitude, speed, distance, direction, origin, and destination. Some of this information changes continually; other pieces of the information are relatively static. Since controllers in a large metropolitan airport may have to keep tabs on a hundred or more aircraft at any one time, you can readily see that such a system must be capable of handling a very considerable amount of information

Another familiar example of modern systems that deal primarily with information are the giant computers that are becoming almost as common as desk calculators and slide rules were a few decades ago.

An essential part of many of these information-handling systems is that they must portray, or display, the information

they are processing so that their human counterparts can see it and interpret it quickly and correctly. When large amounts of information have to be displayed for human users, color turns out to have some extremely useful properties (see, for example, Harris et al.,[13] Morgan et al.,[14] Smith,[15] and Smith and Thomas[16]). Because colors can be so effective for coding information, the engineer or systems designer sometimes wants to know what is the maximum number of different colors that can be used for this purpose. This is quite a complex question because, to be useful for coding information, a set of colors must satisfy some very special requirements. First, any color in the set should never be confused with any other color in the set. Second, every color in the set should be readily associated with a common color name. Third, some color codes should be usable by ordinary people with little or no specialized training in color. And therein lies the origin of our problem for today.

The Dimensions of Color Space

Color sensations can be classified and ordered without any reference to the characteristics of the stimuli that arouse them. Over the past several hundred years, artists, philosophers, and scientists have devised many such classifications of colors according to their similarities and differences. Almost without exception, all such schemes agree that some sort of a three-dimensional model is needed to represent adequately the full gamut of color sensations which the normal person experiences. Figure 5 shows a diagram which is widely used by psychologists and color scientists for this purpose.

Hue. Hue is perhaps the most important of the three fundamental variables of color as a mental phenomenon. It is the main *quality* factor in color and is what the ordinary person means when he says color. Another way of saying it is that hue is the essential element that leads us to refer to colors by such

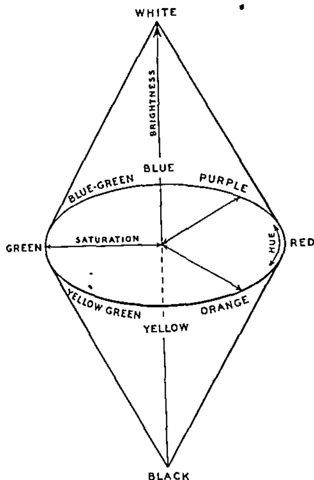


FIG 5 A schematic representation of the psychological dimensions of color space

distinctive names as red, yellow, green, blue, violet, and so on. Hue sensations do not, however, occur in discrete groups Instead, they shade imperceptibly from one to another and, indeed, form a complete circle, as illustrated in Figure 5 Starting with red, for example, one can describe color sensations which become progressively yellower, that is, the red

first becomes orange-red, then orange, yellow-orange, and finally yellow. From yellow one can proceed similarly by continuous gradations to green, blue, violet, and back to red again.

Brightness (lightness). The second variable of color sensation is brightness, the *quantitative* aspect of color sensation. It is easy to imagine two colors, say red, of identical hue but differing in brightness. Common terms which are used to refer to variations in brightness are *light* and *dark*. As with hue, however, the brightness dimension also forms a continuum, shading imperceptibly from very light to very dark hues. In technical work, *brightness* is used to refer to variations in the intensity of lights; *lightness* to variations in the intensity of surface colors.

Saturation. The third variable of color sensation is the most difficult to explain in words alone, without reference to actual color samples. Perhaps the best way of defining saturation is to say that it is the percentage of pure hue in a color. In this sense, it is roughly parallel to the concept of the purity of a chemical compound. In common speech, words such as *pale* or *deep*, *weak* or *strong* appear to refer to variations in saturation. Light brown, for example, is a weakly saturated yellow red of medium lightness, and moderate pink is a weakly saturated light red. In Figure 5, saturation is represented by radii originating at the center of the diagram and extending in all directions from the center. In this diagram, white, gray, and black are colors of zero saturation and thus of no hue. The white–gray–black continuum varies only in brightness, or lightness.

Color space defined. Now, when I say *color space* I mean that three-dimensional space which models color in all possible variations of hue, lightness, and saturation.

The Munsell color system. Although one can study color sensations without reference to physical stimuli, it is nonetheless

true that sensations of color are most consistently and readily elicited by stimuli of the appropriate kinds. The most useful general set of surface colors available for scientific and technical work is found in the Munsell *Book of Color*. This is a large collection of carefully prepared colored chips, spaced to represent about equally noticeable color differences Taken together, the Munsell chips cover virtually the full gamut of all colors reproducible by ordinary paints and pigments. The master colors of the Munsell system have also been deposited with the National Bureau of Standards, where they have been measured and calibrated so that they can be related to other color measurement systems In addition to illustrating the dimensions of color space, the Munsell system is extremely useful for specifying colors in business, science, and industry If a manufacturer specifies the color of one of his products in Munsell terms, anyone else can go to a Munsell atlas and find out precisely what color he means

For all its advantages, the Munsell system is a collection of standardized colors useful primarily as a system of color nomenclature and specification for scientific and technical work The language of the Munsell system is foreign [illegible] sounding [illegible] where the 2 5 YR refers to a particular hue, the 4 to a specific lightness (or value), and the 6 to a particular saturation (or *chroma*) One can easily find out with great precision exactly what such a color looks like, but the terminology of the Munsell system is hardly one that could be used for ordinary color descriptions

I shall refer to the Munsell system later when I come to some of the experimental work I want to discuss At that time, however, my interest will not be in the Munsell system as a system, but only as a comprehensive set of stimuli for [illegible]ing color names

On the Total Number of Discriminable Colors

One approach to our problem of colors and color names is to find out how many discriminable colors there are in color space. No one really knows this. We can, however, arrive at an estimate of this number[17] by discovering how many discriminable colors there are in the hue circle (about 200), multiplying these by the number of discriminable steps in the lightness dimension (about 450), and multiplying that product by the number of discriminable steps in saturation (this varies from about 15 to 165 in different parts of color space). The final outcome of such a series of computations is a very large number—about 7,295,000 discriminable colors!

Although this might appear to constitute an enormously useful reservoir of colors upon which to draw, it turns out to be of virtually no use at all for the kind of practical problem we started with. Let's look at the difficulties with this approach for, in so doing, we will be able to see more clearly the complexities of the problem.

Absolute versus comparative judgments. The first difficulty with the computations above is that they are based on values obtained with the most sensitive methods available to the psychophysicist. Generally, this means two colors presented side by side, under good viewing conditions, and with the observer being required to state merely whether the two are the same or different. This involves a comparative judgment, a kind of discrimination which the human eye can make with exquisite precision.

But the situations in which colors are used for coding information do not permit such niceties of observation. In an information display, a color may appear by itself and the observer may be required to identify it as being this or that. Observations performed under these conditions are sometimes referred to as *absolute judgments* in the psychological literature.[18]

One of the fascinating things about these two kinds of situations is the tremendous shrinkage that occurs when one goes from comparative to absolute judgments. Take, for example, the visible portion of the electromagnetic spectrum, that region stretching from approximately 380 to 780 $m\mu$. If one explores this region step by step, using the most refined methods available to the psychophysicist, one will discover that there are upward of 150 discriminable wavelengths in the visible spectrum.[19] Now present a set of spectrum colors one by one and ask the observer to attach an arbitrary label to each. Even after observers have had extended practice, the maximum number of different colors which they can identify without appreciable error is no more than 12 or 13.[18, 20] This shrinkage from 150 to less than one tenth that number represents a tremendous contraction in the pool of colors available for practical work.

Individual differences. There is still a further complication that we must face in this business if we insist that colors be labeled with common names instead of arbitrary labels. This complication is that people appear to differ in what they see—just as they differ in height, weight, reaction time, or in any of the thousand or more ways we can measure them. Incidentally, let me make it absolutely clear that in this discussion I am eliminating from consideration all persons with color-vision defects. The differences I am concerned with are differences one can find among so-called color-normal individuals.

Let us return again to our visible spectrum, spread it out before an observer, and ask him to locate that part of the spectrum that represents the purest orange, the purest yellow, the purest green, and so on. Let the observer take his time and ask him to repeat his determinations several times. Compare the results you get from several observers and you are sure to discover a remarkable thing. What appears to one observer as a pure yellow looks orange to another. What one

observer is willing to call a yellowish-green, another will call bluish-green. And so on. Each observer is internally consistent, that is, he can repeat his own determinations with remarkable accuracy. But the differences between individuals are disconcertingly large (see, for example, Dimmick and Hubbard,[9] Halsey,[12] and Pickford[10]).

What is the source of such inconsistencies? Do they arise from genuine differences between visual systems – differences between the way your eye and mine respond to the same stimuli? Or are such discrepancies primarily due to differences in the way we have learned to attach color names to our color sensations? These are intriguing questions and, at the moment, we can only speculate about the answers to them. But, however they arise, differences in the way each of us uses color names are one of the added complications we shall clearly have to face in our search for color names that we can use in color space.

How Many Color Names Are There?

Another way of tackling our problem might be to go at it by way of color names themselves. What is the situation here? No one really knows how many different color names there are. Kelly and Judd prepared a dictionary of color terms for the National Bureau of Standards in 1955, and that compilation contains over 5,000 different color names. The difficulty, however, is that the number of color names, unlike the number of discriminable colors, is not some fixed, finite number. There are people in our world of advertising and industry who dedicate themselves to the invention of new color names. Indeed, the guiding principle for people in this kind of work is "New names for old colors every year." As a result, we find ourselves being deluged with such exotic names as *afterglow, air castle blue, Aladdin's lamp, Andrinople berries, angel blue, angel red, apache, aphrodite, April sky, Arab, arabesque, atlantis, atone-*

ment, Australian pine, autumn blonde, autumn brown, autumn glory, autumn gold, autumn green, autumn leaf, autumn oak, and *autumn tan,* just to name a few that begin with the letter "a." Rest assured that these are not my inventions. Some manufacturer, somewhere, has used each of these names to refer to some *particular shade of lipstick, face powder, fabric, or tile.*

There isn't even agreement in the different segments of industry about the precise definition of some of these terms. For example, in the Plochere color system used by interior decorators, *autumn brown* refers to a strong yellowish brown, a yellowish brown of high saturation. In the Standard Color Card put out by the Textile Color Card Association, however, *autumn brown* refers to a moderate brown, a brown very low in saturation

If we eliminate all obvious inventions of the advertising arts, we can still find many color terms which have little meaning for the average person Indeed, it is extremely unlikely that terms such as *alabaster, amethyst, caeruleum, carmine, chartreuse, claret, cyan, ebony, fuchsia, heather, heliotrope, hemp, indigo, japonica, jasmine, madder,* and *mauve* can even be found within the vocabulary of the average American adult. Let me remind you that, according to the American Education Association, in 1960 the median adult in the United States had completed only eleven years of schooling The very fact that you are reading this means that you are a member of a select minority of our population Once again, when we find people who know and use specialized color terms in their work, we often find disagreements *in what they mean by these words* In the field of color photography, for example, *cyan* refers to a color between blue and green To an artist, however, *cyan* is a much bluer color What an artist calls *red* is *red magenta* to a color photographer And so on.

I think you will agree that our brief excursion into the world of color names has so far been a singularly unrewarding exercise One quickly becomes bewildered by their sheer numbers,

confused by the lack of agreement about what they mean, and mystified by their exotic and esoteric qualities.

Color names in common use. Let us now follow another avenue and ask what color names one finds in common use. One of the few studies I know of this problem is reported by Pauline Evans.[21] She surveyed seventeen best-selling novels and found 4,416 color terms used in these books. Of that total number, 4,081, or 92.4 per cent, were accounted for by only twelve terms. These, arranged in order of most to least common, are: white, black, blue, red, gray, green, brown, gold, yellow, pink, silver, and purple. Of the total of 4,416 terms, 4,066, or 92.1 per cent, were unmodified. Only 7.9 per cent were modified with words like "dark," "pale," "pinkish," and so on.

The common language, in other words, appears to get along with no more than about a dozen color names. But surely there must be some middle ground. We can't help feeling that if we look hard enough we ought to be able to find more than a dozen color names to blanket the vast domain of color we can all see.

The Denotative Meaning of Some Common Color Names

My latest approach to this problem has been a direct frontal assault on the question of what people mean by various color names. In particular, I have been looking for consistencies, and for differences, among color names in terms of how normal people use them in designating real colors.

There are at least two ways one could go about this kind of problem. One is to show an observer a large number of different colors and ask him to name them. This technique, I think, has a number of disadvantages. If you do not limit the number of color terms you allow the observer to use, you will generally end up with such a large assortment of different names, with and without qualifiers, that it is difficult to know

what to do with them. There is no easy way of quantifying the outcome of an experiment of that type.

For this reason, we did the reverse kind of experiment. We spread out a large array of colors before an observer, then gave him a color name and asked him to find the one color which, in his opinion, best exemplified that color name. In short, we asked the observer to define what each color name meant to him, by pointing to a particular color.

The color names tested. Table 2 shows the color names we have tested. As you see, the color names are made up of two parts, what we call (1) basic color names, and (2) modifiers. By and large, these basic color names and modifiers were selected from the National Bureau of Standards dictionary of

TABLE 2. THE BASIC COLOR NAMES AND MODIFIERS USED IN THIS STUDY. THE GROUPINGS OF THE BASIC COLOR NAMES SHOW THE SIX SETS INTO WHICH THE COLOR NAMES WERE DIVIDED DURING THE ADMINISTRATION OF THE TESTS. EXCEPT AS NOTED BELOW, EVERY COLOR NAME WAS USED WITH EVERY MODIFIER

Basic Color Names		*Modifiers*
Red	Green	(No modifier)
Pink§	Yellowish green	Vivid
White*	Yellow green	Strong
Gray†	Greenish yellow	Pure
Black*	Olive	Deep
		Dark
Yellow	Blue	Light
Orange	Bluish-green	Pale
Brown	Greenish-blue	Grayish
		(ish) White
Purple	Reddish-purple	(ish) Gray
Violet	Purplish-red	(ish) Black
Purplish-blue	Purplish-pink§	

* No modifiers were used with these color names.

† This name was used unmodified and with each of the following modifiers: *Light, Dark, Pure, Medium.*

§ The modifier -ish *Black* was not used with these color names.

color terms,[22] although we have introduced some additional terms of our own. With the few exceptions in the footnotes to the table all basic color names were paired with all modifiers. Some examples of our color names are *vivid green*, *pale blue*, *deep olive*, *grayish-purplish blue*, and *violet black*. These pairings of modifiers with names gave us a total of 233 color names, the number tested in this investigation.

General testing procedure. For our colors we used all the color chips in the Cabinet edition of the Munsell *Book of Color* (1,226 in all) and added to them 133 special colored chips generally of higher saturation than are contained in the book. We ended up, therefore, with a total of 1,359 different samples of color.

These colored chips were spread out on a large table, under a carefully controlled source of illumination approximating natural daylight. The observer stood, or, if he wished, sat, before this display. He was given a color name, for example, *vivid brown*, and instructed to find that one color which, in his opinion, was the best vivid brown on the table. He was given unlimited time to make his selection and he was allowed to pick up the cards containing the colored samples, move them about, or handle them in any way that would make his search easier.

To reduce this task to manageable proportions, we did the testing by groups of colors (as shown in Table 2). For example, all of the color names containing *yellow*, *orange*, and *brown* were tested as a group. In this way we were able to modify the size of the display and the number of colors through which the observer had to search. We made sure, of course, that the range of colors displayed was more than adequate to cover those he might want to search through. At any one time, however, a display contained no more than about 550 colors.

The six different sets of colors, and all the color names within a set, were tested in a different random order for each of the observers used.

The observers. The observers were forty young adults, twenty

men and twenty women. They were required to have normal color vision, good visual acuity, and to speak English as their native tongue. The observers were housewives, students, secretaries, and teachers. As a group they were undoubtedly above the average intelligence of the population as a whole.

The basic form of the data At the conclusion of many weeks of testing we ended with 9,320 color selections—10 for each of the 233 color names. Table 3 shows the way in which the raw data were tabulated for four of them. It also shows some of the statistical measures that were computed to describe the selections.

Our first interest was in getting some measure of the consistency with which the forty subjects could agree in their selections. To this end we first determined n, the total number of different selections made for each color name. Next we computed the frequency of the modal, or most common selection, M, the sum of the frequencies of the two most common selections, M', and the sum of the frequencies of the three most common selections, M''. Finally, we used a rather complicated statistic which I borrowed from communication theory, although we used the measure simply as a summarizing statistic having nothing to do with communication. This measure, which I call the coefficient of consistency, is defined by

$$C = 1 - \frac{U_{sel}}{U_{max}} = \frac{[\Sigma(f \log_2 f)]}{[40 \log_2 10]}$$

where U_{sel} is the uncertainty, or amount of information among the selections made by all forty observers, U_{max} is the maximum uncertainty, or maximum amount of information, possible among the selections, and f is the number of times each of the n color chips was selected. This measure of consistency, or agreement, ranges from 0.00 to 1.00. If all forty subjects made exactly the same choice for a color name, the value of C would be 1.00. If each of the forty subjects made a different choice for a color name, the coefficient would be 0.00.

observers each agreeing on one of thirteen colors and the fortieth picking a fourteenth color. A value of 0.52 could occur if the forty observers had picked only six different colors—seven observers each agreeing on one of five colors with the remaining five agreeing on the sixth. To sum up, then, the agreement among color selections is better than the distribution in Figure 6 seems to suggest.

Consistency of selections made by the two sexes. Since we tested twenty men and twenty women, it is natural to ask: Were there any sex differences, and, if so, which sex was more consistent? The answer is that the women were significantly more consistent. This finding is in agreement with those of other studies which show that women or girls tend to be better than men or boys in other kinds of color naming (see, for example, DuBois,[23] and Ligon[24]).

Consistencies for groups of color names. Figure 7 reveals some

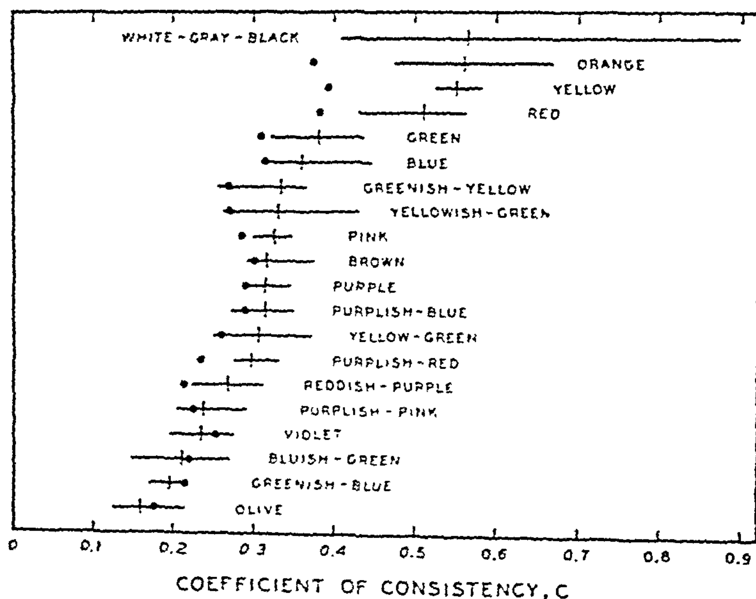


FIG. 7. Consistencies of selections made to the basic color names.

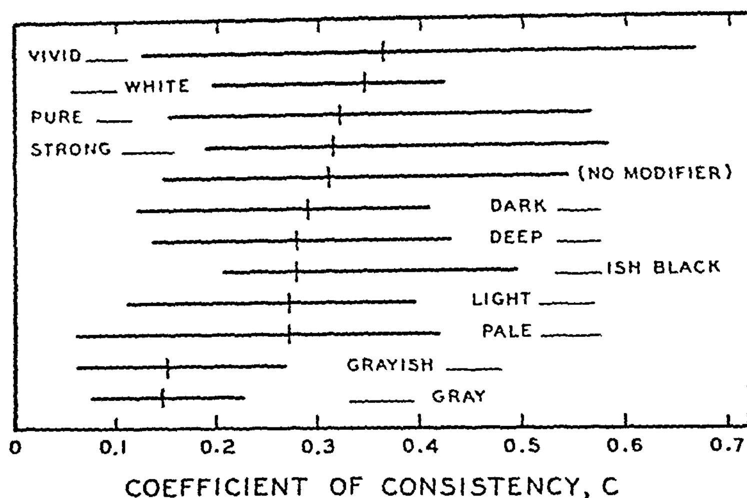


FIG. 8. Consistencies of selections according to color name modifiers.

name (see Table 3 for examples). I shall discuss here only some of the findings for the hue selections.

The mean hues for all the *strong* colors are shown in Figure 9 as arrows radiating out from the center. The scale and the symbols around the inside of the circle show the locations of the forty principal Munsell hues. The short arcs on the outside of the circle are ranges for *strong* hues as defined by color experts of the Inter-Society Color Council and the National Bureau of Standards.[22] In some cases, the agreement between the ISCC–NBS definitions and what our observers did is excellent. This is true, for example, of *strong red*, *strong orange*, and *strong yellow*. In other cases, the agreement is poor. Note, for example, that the mean selection for *strong green* is completely outside the range of values specified by the ISCC–NBS. A particularly interesting comparison is that between *strong purple* and *strong violet*. Color experts agree that violet should be a bluer color than purple. Our observers, however, could make practically no distinction

Fig. 9. Average hue selections for the *strong* hues (arrows). The scale and symbols around the inside of the circle are designations for the Munsell hues. The segments of arcs on the outside of the circle show the ranges of hues defined by the ISCC–NBS method of designating colors.

between these two colors. In fact, they even reversed the two!

Another highly interesting point concerns the compound colors, such as *greenish-yellow*, *yellow-green*, and *yellowish-green*. These turn out to be particularly difficult. Although color experts think of *yellowish-green* and *greenish-yellow* (or *bluish-green* and *greenish-blue*) as being distinctly different colors, our observers scarcely differentiated them at all.

The last point of interest in Figure 9 concerns the distribution of the mean selections. Recall that the Munsell hues

are spaced to correspond to equally noticeable color differences; that is, equal distances around the perimeter of the circle correspond to equally discriminable hue differences. When we first made our selection of nineteen basic color names for this study (those in Table 2) we tried to pick names that would blanket the hue circle almost uniformly. Our findings show that we fell far short of this goal. There is a large region between most of the Munsell greens and all the blue-greens that was not sampled at all by our color names. There is another large empty space between *greenish-blue* and *blue* and still another between *violet* and *purplish-pink*. These regions contain a great many colors which the eye can discriminate if it wants to. But apparently we have not found these colors to be sufficiently interesting to reward them with their own distinctive names.

On the overlap between different color names. The selections of color chips were sometimes nearly identical for different color names. Table 3 illustrates some of these. Note, for example, that 5.0 R 4/14 was the most common selection for *red*, *pure red*, *vivid red*, and *strong red*. In fact, only two Munsell chips (5.0 R 4/14 and 7.5 R 4/14) accounted for over half of all the color selections made for these four color names. Another way of saying this is that our observers regard *red, pure red*, *vivid red*, and *strong red* as being essentially the same color. This conclusion is supported by the means of the selections made to these four names (Table 3). The maximum discrepancy among the four mean hues, mean values, and mean chromas is less than half the difference between any two color chips adjacent in the Munsell system.

Two measures were computed to quantify the overlap between the selections made to pairs of colors. One measure was a simple tally of the number of times identical chips were selected for different color names. The other, again borrowed from communication theory, is a measure of the information transmitted (or, as it has also been termed, the contingent

uncertainty[25]) between color names and color selections. Here again, think of this only as a statistical measure of association, varying, for these data, between 0.00 and 1.00. These two measures agree so highly that only the first was computed for all the data. Measures of overlap are now available for all 27,028 pairs of colors so that we can determine which color names overlap and by how much.

The results of this analysis reveal that there is almost complete synonymy between *violet* and *purple*, and between *yellowish-green* and *yellow-green* This identity holds not only for the basic color names but for the basic color names used with all eleven modifiers Although there is somewhat less overlap between *yellow-green* and *greenish-yellow*, between *bluish-green* and *greenish-blue*, and between *reddish-purple* and *purplish-red*, it is sufficient for us to conclude that we could not expect people to make dependable discriminations between the members of each pair

The situation with regard to modifiers is, in some respects, even more interesting First, the modifiers *pure*, *strong*, and *vivid* appear to add nothing to a basic hue name. For all practical purposes, *pure blue*, *vivid blue*, and *strong blue* are synonymous with just plain *blue* Deep and *dark* also turn out to be synonymous, as do *pale* and *light* Moreover, these generalizations hold for all nineteen basic color names. All these findings are at variance with the ISCC–NBS system of designations in which these modifiers were supposed to identify distinct variations in lightness and saturation Unfortunately the ordinary person does not make these distinctions

There are also numerous other pairs of color names which overlap, but these are more or less isolated instances and do not hold for entire sets of modifiers or basic color names For example, although the *pink* and *red* colors appear to be differentiated on the whole, *pinkish-white* and *reddish-white* are not Similarly, *orange* and *brown* appear to be distinctive except for *orange-black* and *brownish-black*

On the total number of distinctive color names. We turn now to what is probably the most interesting question, essentially the question we started with: How many distinctive color names are there? Imagine that we were to construct a very large table of data. Across the top of the table we list the color chips, starting with the first and continuing through the 1,359th. Along the side of the table we list color names, starting with the first and ending with the 233rd. Inside the table we enter in the appropriate cells all the 9,320 color selections made by our observers. What we will have, in short, is a kind of gigantic correlation table, containing 1,359 times 233 cells.

The measure of information transmitted (or contingent uncertainty), to which I have already referred, is a kind of statistic which is ideally suited for data of this type. If we calculate such a measure for this large table of data we obtain a result equal to 5.49 bits. The antilog of 5.49 bits is 45. This means that there are theoretically 45 different color names for which color selections should not overlap.[25]

Unfortunately, when we examine our table of data we find that, in still another respect, our selection of color names was not as good as we had hoped. Of the total of 1,359 chips, 246, or about 18 per cent, were never selected at all. Our theoretical number of 45 color names does not cover the entire color space of the Munsell system but only about 82 per cent of it.

We can, however, make some estimate of what the total number of color names could be and bracket it with some upper and lower limits. For the lower limit, let's assume that we managed to find some additional color names to sample the unused 246 color chips. Let's assume, further, that the agreement for these new color names is no better than the agreement for that of our most variable color name, *pale olive.* Under these assumptions we would need an additional seven color names to cover the unsampled area of the Munsell system.

Let's follow another line of reasoning to see what it will give

us as an estimate. If 45 different color names can cover 1,113 chips, on the average, how many color names would we need to cover 1,359 chips? That turns out to be about 55.

By putting all these computations together, we may say that the total number of different color names that could be used for color space (as represented by the Munsell system) is most probably between 52 and 55.

Postscript and Prospect

Let's stop now, take a look backward at what I've tried to say, and then a look forward at what this means for future research.

I started with a question: What is the maximum number of usable color names for all of color space? We saw first that the normal human eye is capable of discriminating upward of 7,000,000 different colors, but that this number is not at all a fair estimate of the reservoir of practically usable colors. We saw then that the English language contains thousands of different color names and that our language is flexible enough, and our culture rewarding enough, to encourage the proliferation of color names, although most of us have little idea what these color names mean. On the other hand, the common English language, the language of the novel, seems to get along well with no more than about a dozen different color names.

A direct experimental attack on the denotative meaning of various color names has revealed some interesting things about what people mean by various color names. We saw that some color names which appear to be quite different are, for all practical purposes, synonymous. We saw also that people do not make distinctions between certain modifiers which appear to convey the idea of variations in saturation and lightness. And, finally, we ended with the estimate that there are probably between 52 and 55 usable color names for all of color space.

As is so often the case with scientific research, the questions we have uncovered are far more numerous than those we have answered. Our 52 to 55 usable color names are theoretical, rather than actual. Inspection of our data identifies a number of names which do in fact refer to different regions of color space. But what of the unsampled parts of color space? Can we find color names for them? Are there indeed simple color names which we can use for the hues that lie between green and blue-green? Or for the low-saturation hues, which were also conspicuously avoided?

How do we learn color names? Where and how in the genetic development of a child are associations established between color names and color sensations? How does the child learn to make the discriminations that provide him with the color language he has as an adult?

Would we have obtained the same results if we had tested observers who know British English instead of American English? Would a French psychologist have discovered the same number of usable color names as we did? Or a German psychologist? In short, have we discovered something reasonably constant about human color perception, or have we discovered something that is specific to people who speak a particular language?

These are only a few of the intriguing questions that occur to me, and I am sure, to you as well.

In conclusion, if my brief remarks have colored your thoughts with the warm glow of *crushed strawberry, sunglow, coral blush,* or *mountain haze*, I will feel that I have, in part at least, earned the privilege of addressing you. But if, by some stroke of good fortune, I may have enticed a few of you into exploring these questions further, I will feel that I have been most richly rewarded.

I am grateful for the assistance of Dr. Aleeza C. Beare, who collaborated with me in some of the research described here.

The work reported in this paper was done in part under Contract Nonr-248(55) and in part under Contract Nonr-4010(03) between the Office of Naval Research and The Johns Hopkins University. It is Report No. 12 under the latter contract. Reproduction in whole or in part is permitted for any purpose of the United States Government.

REFERENCES

1. M. RICHTER. *Internationale Bibliographie der Farbenlehre und ihrer Grenzgebiete. Nr. 1: Berichtszeit 1940–1949.* Musterschmidt, Göttingen (1952).
2. M. RICHTER. *ibid., Nr. 2: Berichtszeit 1950–1954.* Musterschmidt, Göttingen (1963).
3. J. M. CATTELL. Ueber die Zeit der Erkennung und Benennung von Schriftzeichen, Bildern und Farben. *Philosophische Studien* 2, 635 (1885).
4. R. F. BERDIE. Effect of Benzedrine Sulphate on [illegible] Color Naming. *J. Exp. Psychol.* 27, 325 (19[illegible]
5. J. M. CATTELL. Mental Tests and [illegible] 373 (1890).
6. J. M. CATTELL and L. FARRAND. [illegible] ments of the Students of [illegible] 3 618 (1896).
7. F. REMOLI. Farbenkenntnis und [illegible] schülern. *Z. pädagogische* [illegible] (1933).
8. A. C. BEARE. Color-name as a [illegible] *J. Psychol.* [illegible] 245 (1963).
9. F. L. DIMMICK and M. R. H[illegible] [illegible] Psychologically Unique Yellow [illegible] 52 242 (193[illegible]
10. R. W. P[illegible] [illegible] and Kegan Paul, London [illegible]
11. D. KATZ and [illegible] Normal and Anomalous [illegible] (1934).
12. R. M. HALSEY. [illegible] White and Purple. *J. Opt. Soc.* [illegible]

13. D. H. Harris, P. E. Kolesnik, and K. S. Teel, Wire Sorting Performance with Color and Number Coded Wires, *Human Factors*, 6, 127 (1964).
14. C. T. Morgan, J. S. Cook, III, A. Chapanis, and M. W. Lund (Eds.), *Human Engineering Guide to Equipment Design*, McGraw-Hill, New York (1963).
15. S. L. Smith, Color Coding and Visual Search, *J. Exp. Psychol.*, *64*, 434 (1962).
16. S. L. Smith and D. W. Thomas, Color Versus Shape Coding in Information Displays, *J. Appl. Psychol.*, *48*, 137 (1964).
17. D. Nickerson and S. M. Newhall, A Psychological Color Solid, *J. Opt. Soc. Am.*, *33*, 419 (1943).
18. A. Chapanis and R. M. Halsey, Absolute Judgments of Spectrum Colors, *J. Psychol.*, *42*, 99 (1956).
19. W. D. Wright, *Researches on Normal and Defective Colour Vision*, Mosby, St. Louis (1947).
20. R. M. Halsey and A. Chapanis, On the Number of Absolutely Identifiable Spectral Hues, *J. Opt. Soc. Am.*, *41*, 1057 (1951).
21. R. M. Evans, *An Introduction to Color*, Wiley, New York (1948).
22. K. L. Kelly and D. B. Judd, *The ISCC–NBS Method of Designating Colors and a Dictionary of Color Names*, National Bureau of Standards Circular 553. Washington D. C., U. S. Government Printing Office (1955).
23. P. H. DuBois, The Sex Difference on the Color-Naming Test, *Am. J. Psychol.*, *52*, 380, (1939).
24. E. M. Ligon, A Genetic Study of Color Naming and Word Reading, *Am. J. Psychol.*, *44*, 103 (1932).
25. W. R. Garner, *Uncertainty and Structure as Psychological Concepts*, Wiley, New York (1962).

6. THE FINE STRUCTURE OF THE BACTERIAL CELL AND THE POSSIBILITY OF ITS ARTIFICIAL SYNTHESIS

By Ernest C. Pollard
Pennsylvania State University

In the past decade a scientific revolution has taken place, which is perhaps the greatest scientific revolution of all I refer to the achievements of a strangely assorted group of geneticists, biochemists, virologists, physicists, classical biologists, and even engineers, who have founded the strangely named, but powerful, subject of molecular biology. These individuals are not a homogeneous group, they are bound by little more than an intense intellectual fervor they can disagree sharply with one another, and they accord respect to any one with the greatest personal reluctance These are all symptoms of men of science who have suddenly developed sharp insight and who are building a doctrine which will be used to interpret living things for the future, and probably the long future

To set this revolution in perspective, let us look at a comparable revolution in physical science which took place at the turn of the twentieth century, when in a few years we witnessed the discovery of the electron, of X rays, of radioactivity and the atomic nucleus, of the interpretation of spectra, and of the nature of solid matter In about forty years from the

first of these discoveries the claim could fairly be made that a complete understanding of normal inanimate matter had been attained by science. In this claim cosmology is excluded, and also the high-energy features of the atomic nucleus. What is meant is essentially this: that for anything which can be touched or perceived in the ordinary room, one can find, somewhere in the world, an expert who can tell you the precise structure of that object, how to alter it to your will, or that it cannot be so altered if it is inherently impossible to do so. Some latitude must be allowed in this claim, bearing in mind that expense and time could be involved, and we are cheerfully assuming that they represent no problem.

Now the suggestion I am making about the revolution in biology is that in thirty more years, or, again, forty years from the start, it will be possible to name any part of an organism, to find someone somewhere in the world who can tell you its precise structure, how to alter it to your will, or that it is inherently impossible to do so. This revolution, unlike the revolution that concerned inanimate things, not only affects our environment, making it easier for us to live in it, but it concerns ourselves as living beings. If we widen our imagination a little, we can see that we shall not only have a profound ability to control disease, far beyond our present ability, but we shall also be able to control factors which now are outside our powers. To show the kind of potential ability we might have, we should by then understand memory, and indeed intelligence, and we should therefore be able to cure our memory and intelligence defects by going to the appropriate biological engineer and getting the proper operation done on us. Clearly, the social revolution that could result is not to be belittled, and it is indeed latent in the growing power of interpretation given us by molecular biologists.

If this claim is made honestly, then it should be followed logically to the conclusion that if we understand a cell we should be able to make one; hence the title of this article and

the examination of what the task would entail, and what might be the discoveries made on the way, if any. So my goal is to try to present the character of the structure and working of the bacterial cell in a way that will provoke thought about the problem of its artificial synthesis.

Three further things need to be said before this task is undertaken. The first concerns the choice of the bacterial cell. Why not a virus? Why not a human cell?

A virus has been, to my mind, already synthesized. There are at least four laboratories in the United States alone where nucleic acids are being made from cell-free systems One of these nucleic acids has presumably had the necessary length and complexity to provide the code for two enzymes and so one can say that a virus has been made. The problem is finding the necessary host, and this highlights the basic objection to choosing a virus to represent the achievement of construction of a whole cell it is not, alone, fully representative of life. So, at least as far as I am concerned, I cannot feel that the true challenge has been met when we have made a virus. Now if we choose to set up the task of making a human cell, we are involved with something that is not normally found out of contact with 10^{12} other cells in some living being, with a cell that bears the characteristics of that sheltered existence, unusual size, and the presence of subunits which are part of its functioning—organelles such as mitochondria, for example. At the moment, it seems to be unnecessarily hard, though biological discovery may well show this conclusion to be wrong one day However, for the present let us accept this view The bacterial cell, which is quite capable of autonomous existence, under conditions which are most adverse, is certainly living[1] and certainly a challenge sufficiently great to us at present

1 Studies made by individuals interested in means for testing for life on Mars[1] indicate that nowhere on the surface of the earth can one select a gram of soil without also selecting 10,000 (minimum) living organisms

The second preliminary thought concerns what we mean by the artificial synthesis of a cell. By this I mean that we could start from the elements of a cell: carbon, nitrogen, hydrogen, oxygen, phosphorus, and sulfur, plus the trace factors, and assure ourselves that there was no mystery about completing each necessary step all the way from the elements to the newly living cell made from those elements, without necessarily actually completing more than one step at a time. If some necessity arose to marshall money and scientists to do the whole task, as was found to be the case for controlled nuclear reactions and radar, then it could be done. *Knowing how to do it* is what we are discussing.

The third preliminary thought is of great importance, especially to biologists. To make the point, I shall return to modern physics. The discoveries concerning the atom and the atomic nucleus, particularly the latter, a school of thought in which I began my scientific work, could never have been made with our eyes directly; they had to be made by hypothesis and the test of experiment or, in better terms, by *accurate imagination*. Thus the fact that we can never see an atom of sodium doesn't prevent us from knowing the orbitals that characterize its structure, and if we want to construct models that represent our ideas of these orbitals we can do so; and such models are often very useful to students who are interested in the way in which atoms behave, particularly in relation to one another. Molecular biology has already shown that many of the essential operations of the living cell take place at small molecular dimensions, and so we cannot hope to see both the molecule and how it is working at the same time. Thus, we are once again forced to the position of the physical scientist, and must use accurate imagination to describe the cell, if we want to use enough detail to see the mechanisms at work.

This leads to the final introductory remark. In what follows we shall have to draw thought pictures, already known to the

of the Oak Ridge National Laboratory, who started his graduate work in my group at Yale. It is a thin section, and it shows the outline of the cell. The black dots that fill the majority of the cell are probably ribosomes, and the relatively clear part, which shows some evidence of fibrils and some structure in the fibrils, we can call the nucleus. Before we discuss the features of interest in this cell, an inventory of its contents, given in Table 4, is useful to have.

TABLE 4. OUTLINE OF NATURE AND CONTENTS OF A BACTERIAL CELL

Length, 3 microns, diameter, 1 micron
Volume: 2.2×10^{-12} cc
Water 70%: and of what remains:

DNA 3%	Protein 70%:
RNA 12%	Ribosomes (10,000)
	Enzymes
	Surface structural protein
Lipid 6%	
Phospholipid 4%	
Polysaccharides 5%	

DNA in one, two, or three continuous units
Molecular weight 2×10^9
RNA in three forms
Ribosomes: RNA – protein particles, 3×10^6 MW

To comment on this table, we can point out that the cell, at a diameter of one ten-thousandth of a centimeter is already very close to the wavelength of visible light, so that as we look at it in the microscope we already, even for its outline, must consider that it is scattering and diffracting light and judge if our conclusions from what we see are correct. Any internal structure we "see" must definitely be considered as scattering centers and judged accordingly.

The cells appear to be very wet, with 70% water. But once again, remember that the hard-boiled egg at breakfast, with its firm denatured albumin, has 50% water, and isn't very "sloshy." So with that 70%, the bacterial cell hasn't too much

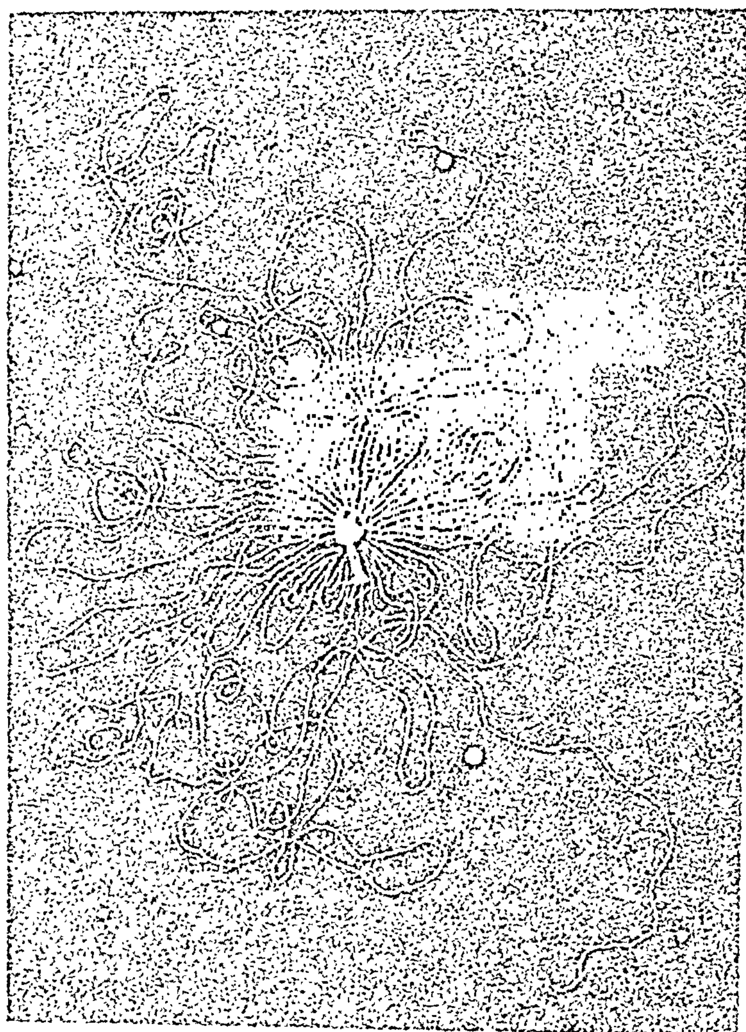


FIG. 11. DNA from a bacterial virus. It can be seen to be a single unit of truly great length. This very spectacular picture is taken from the work of Kleinschmidt et al.[2]

this picture the scale is totally different, and to visualize the bacterium it is necessary to compare it with one of the dots in the emulsion. The line at the bottom is about fifty times the length of the bacterium. The technique used to get this picture is very elegant and is as follows. Advantage is taken of the base *thymine*, which is found only in DNA. It can be readily procured with a tritium label and the tritium, when it undergoes radioactive decay, emits a béta ray of very low energy which will cause the development of only one or two photographic emulsion grains. Thus, the presence of developed grains means the presence of DNA. What Dr. Cairns does is to label the bacterium very fully with this kind of thymine and then with great care and skill, and he is *par excellence* the one who can do this, extract any DNA, place it in contact with a photographic emulsion, and let the image develop as

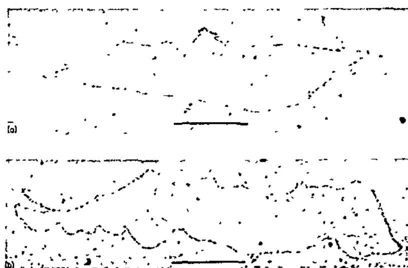
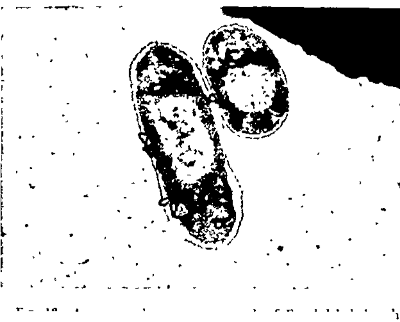


FIG. 12. A radioautograph of the DNA taken from *E. coli* cells. This DNA has been labeled with H^3-thymine and extracted with ... Spring Harbor Labora... *... elation of Ultrastructure,* Cell Biology (Academic Press, 1962)

the tritium decays. The result is what is in the figure, and it shows dramatically how the long continuous piece of DNA has been released from the structural form it must have had in the cell. Clearly, one of our problems is to "put" this DNA inside the cell: it is not simple.

In order to gain some idea of where to put the DNA we can consider some work done by Dr. Caro and Dr. Van Tubergen, working for Dr. Forro at Yale. This work was actually part of a Ph.D. thesis. Again what was done was to label the bacterium with tritium thymine, and in this instance to take random but consecutive slices of the bacteria which had been embedded in methyl methacrylate as if for electron microscopy. The slices were then studied by the same technique of radioautography that we have described, and the question asked whether all slices had about an equally probable chance of developing grains of emulsion, or whether some had no chance at all, because there was no DNA in that part of the cell and so no possibility of any radioactive decay due to thymine. What was found was that considerable numbers of zero grains were present, and by analyzing the data the conclusion was reached that the DNA is confined to a small part of the bacterium which could easily be associated with the part in the middle of the electron micrograph of Figure 10; and indeed, if one looks closely at this quite remarkable picture one can convince oneself that there are fibrils in the relatively empty part, and even that some sort of organization exists there. Thus, we must put the DNA in the nucleus of the cell.

The next question which naturally arises is the location of the RNA. In Figure 13 we show a remarkable combination of electron microscopy and radioautography. In this figure the label has been given as tritiated uracil, which is not found as such in DNA but is mostly in one or another form of RNA. In addition to this change, a very valuable technique of development of the emulsion has been used. Instead of waiting for



side the nuclear region This superb example of technical skill is also due to Dr Lucien Caro

the whole grain of silver bromide to develop, the development is halted quite early and fixation introduced The result is that the inception of the latent image can be seen instead of the fully blackened whole grain, and a far better resolution is so obtained. It can be seen that the origin of the radioautographic images all lie beyond the central part we associate with the nucleus. Therefore, it seems reasonable to assign the RNA to a part of the cell which is outside the nucleus

Inferences Regarding the Cell Structure Preliminaries Regarding Ordered Synthesis

With these figures we really exhaust the part of the description of the cell that can be made reassuring with electron

micrography, and we have to embark on the process of imagining, in a guided way, the kind of structure that must be responsible for the various processes in the cell. To do this we need to have an idea of the most fundamental processes at work, and which must be considered to have their proper place in the cell. It is the character of these fundamental processes which has been elucidated by modern biology, and so we have to give a short account of this subject in order to be able to suggest ways in which the parts of the cell are located and how they work together.

We begin with a compact diagram of four kinds of nucleic acid, shown in Figure 14. On the left is the all-important DNA, which is shown schematically and not realistically. It is in reality in a double helical pattern and so must twist and untwist for some of its functions. For our present purposes the helical structure is not important, and we have simplified the representation. The essential features are the double

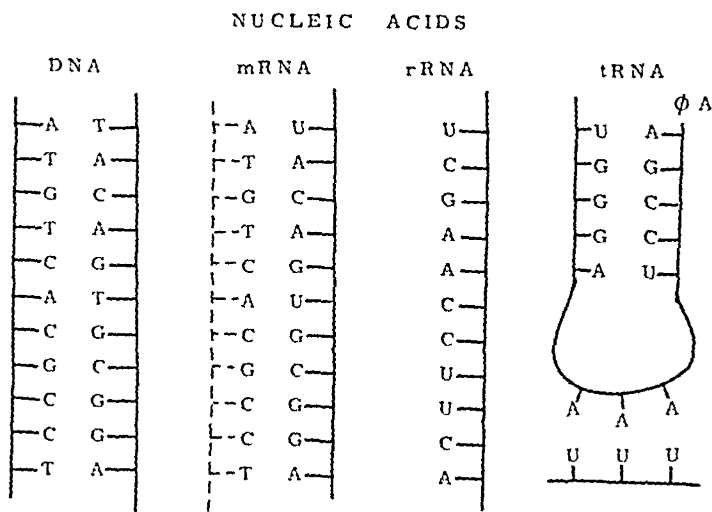


Fig. 14. A very schematic representation of the four kinds of nucleic acid.

structure; both chains are polymers of adenine (A), guanine (G), cytosine (C), and thymine (T), which are linked through a deoxyribose sugar and phosphate and the requirement that, in the chain opposite, every thymine find an adenine and every guanine a cytosine. In this way a preservation of the instructions is made, for, although the two chains are quite different, the existence of only one also guarantees that the other can be made absolutely exactly. DNA also has the remarkable property of being fixed once made. Unlike the other components of the bacterial cell, DNA does not "turn over," but once it is completed remains as such, unless rather drastic external circumstances intervene. This permanence of the DNA is also necessary to make it effective as a hereditary codescript, to repeat the remarkably predictive words of Schrödinger.[4] We need to remember that one million or more bases in a continuous line form the actual molecule of DNA. Again we remind the reader that Schrodinger predicted that this codescript would have to be an "aperiodic crystal." His words do really describe almost exactly the strange kind of structure we see in these giant molecules.

The next kind of nucleic acid is messenger RNA, probably the most exciting and strange material in nature. Intense study of this material will certainly be repaid by great scientific excitement and as advice to a young scientist reading this article, I suggest that he or she settle down to a long occupation with it and its behavior. It is single-stranded, and it is made as a complementary copy of one strand of DNA, which has been indicated in the figure as a dotted strand. This means that the RNA is almost identical with one strand of DNA, and would be identical except for the presence of the base uracil, instead of thymine, and the additional oxygen in the sugar part of the polymer. This messenger RNA has a molecular weight of about one million, making it far smaller than the DNA, though by no means small, and it, together with a considerable apparatus, which we shall shortly

describe, is the means by which the cell guarantees that the instructions on the DNA are used to make protein of the precise kind needed. In bacteria, the messenger RNA is not permanent. It has a half-life which depends on the condition of the cell (temperature and oxygen supply, for example) and which is about one minute at optimum temperature. This means that the amount of protein that can result from one molecule of messenger RNA is limited.

Since we have much more to say about this RNA later, we can consider the other two forms. The ribosomal RNA which is shown next, reading from the left, is, in some ways, not yet understood. It is permanent, in the sense that it does not decay rapidly, and its known function is to form part of the ribonucleoprotein *ribosomes.* It is known that it is also made to be complementary to some stretch of the DNA, but it (the nature of this stretch) is not wholly clear to us. It is relatively bland in its function and because it is harder to study we may have one or two years to wait before we know much more about it.

The last form of RNA is transfer RNA, the little RNA, sometimes called "soluble" RNA. This is exciting at the moment because, just as I began this paper, the remarkable work of Dr. Holley's group at Cornell became known.[5] The features of tRNA that are important to us are first, the provision for attaching a particular amino acid at one end of the RNA, and second, the presence of a triplet "code" of three bases, somewhere about the middle of the molecule, which are capable of establishing a firm relationship with three complementary bases on messenger RNA, and so of bringing the appropriate amino acid into place for incorporation into the chain of a protein. In the schematic diagram, the amino acid selected is phenylalanine, and the code for this is three adenines, which match up with three uracils on the messenger RNA. The tRNA has a secondary structure, which is thought to have a great deal of complementary base pairing within itself, as indicated. Holley's group have shown the entire

sequence of 77 bases, and their work opens up a vast future in the determination of nucleic acid sequences. In their case, the tRNA carried the code for alanine. They show several possible final structures for it, so that even when the whole sequence is known, the whole story is not in.

We need to remind the reader that there are nineteen amino acids, for which the structures are nicely set out in the late Dr. Quastler's very perceptive book, *Information Theory in Biology*.[6] Joining these so that the amino groups and carboxyl groups coalesce with the exclusion of a water molecule causes the formation of the peptide bond, and by doing this to perhaps 100 to 300 such amino acids we obtain the raw material of a protein. For it to be of any biological significance that we know at present, it must have a definite sequence, and this is assured by the "holy trinity" of DNA to RNA to protein.

Ordered Synthesis in the Cell

We now come to the description of one of the absolutely fundamental processes in the cell, which must be described in such a way as to be integrated actually in the cell. To think about it, Figure 15 will help. It can be seen that DNA is at the heart of the process, and that it has two starting functions: one to make more of itself (indicated by the arrow which sweeps upward), and the other to make the general happenings of the cell possible. To do this it must be [illegible] to the "manufacture" of messenger RNA, transfer RNA, and ribosomal RNA as shown. All three of these then [illegible] to form the protein, which is either structural or enzymatic. The enzymatic part handles the supply of materials which come in from the outside and converts them into metabolic use. As a result, the two major "functions" of the DNA [illegible] the "[illegible] phase" are able to function. The cell does [illegible] much more than just what is shown in the figure, but we shall

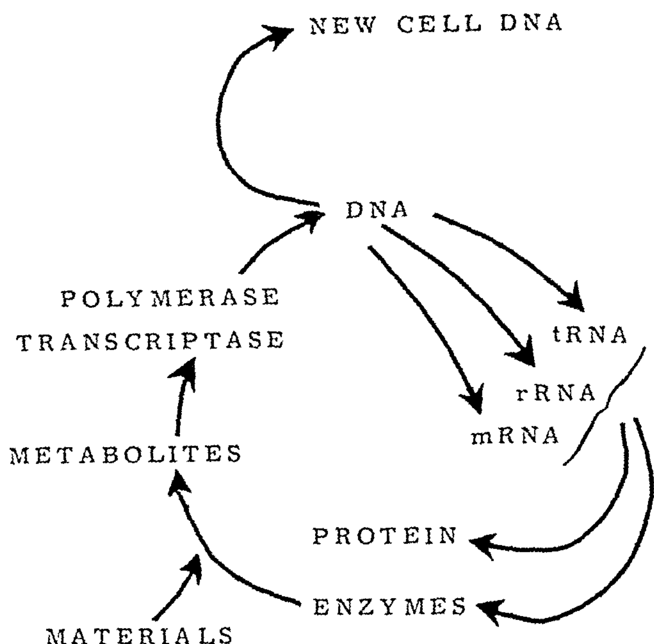


FIG. 15. A schematic diagram of the regulated synthesis of a living cell.

have quite a problem assimilating these into the confines of the tiny volume of the cell as it is.

We can start the process of seeing how these actually must be by looking at the suggestion made by Dr. Cairns as to the way in which DNA replicates (Fig. 16). Since DNA is of helical form, the management of it in any kind of way along its length is bound to require some spinning process. Dr. Cairns suggests that there is a "hypothetical" spinning apparatus which is at one end of the DNA, and we can start at that point. It is shown as the black triangle. The replication of the DNA, which has been shown to progress uniformly in a linear fashion by Yoshikawa and Sueoka[7] and Lark, et al.,[8] is in the

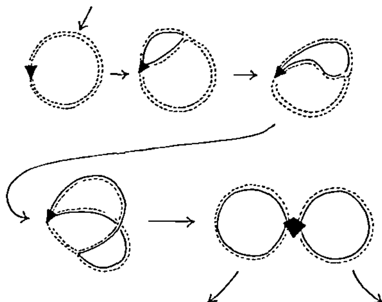


FIG. 16 Dr Cairns' suggestion about the mechanism of DNA synthesis

form of a kind of "Y," and is shown by representing the new *DNA as black* and the *old DNA* as dotted. The synthesis progresses around the continuous DNA strand and, at the end, the spinning mechanism is supposed to replicate also, the two then separate into the DNA needed for the two daughter cells This DNA is both old and new as can be seen

To move on quickly to one of the "factories" of the cell, in Figure 17 is shown a schematic representation of what I believe must be happening The shape need not be in this form, for example, a shape more like a "yo-yo" might be even better or it could be much more spread out, but nevertheless this kind of thing is indicated To comment on it first, the DNA must separate some time, so it is indicated as separating just before it enters the factory The actual point of synthesis is double and takes place where the new strands

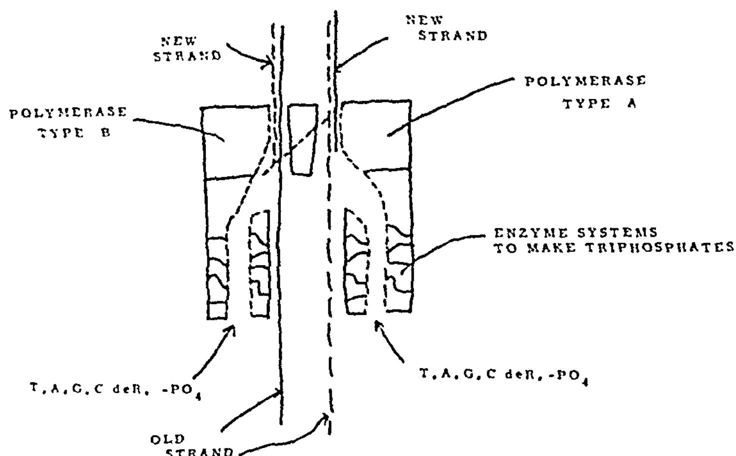


FIG. 17. A schematic diagram of the polymerase "factory," sometimes called the "growing point" for the DNA.

appear. To achieve it I have suggested that two polymerases, each of different specificities, must be present to handle the difficult problem of synthesizing DNA which is presented differently in two strands. The component parts that are immediately needed by the synthetic process are made in situ from the four bases, phosphate and sugar. This requires quite an assembly of enzymes which can make the four components and keep them at high concentration in the little region near the synthetic points.

In reality, the excitement of this kind of figure lies in the rate at which the DNA is moving through the factory. If one takes a pencil and sweeps it up the figure as fast as it can be moved, this is still about ten times too slow. It also has to be realized that the speed is definitely matched by the great accuracy, which is such that only one mistake is made in every 10^8 times when the conditions are at their best.

To justify the need for a "factory," the following piece of reasoning is offered. On Figure 18, the component parts which are assembled into nucleic acid are shown. The TTP,

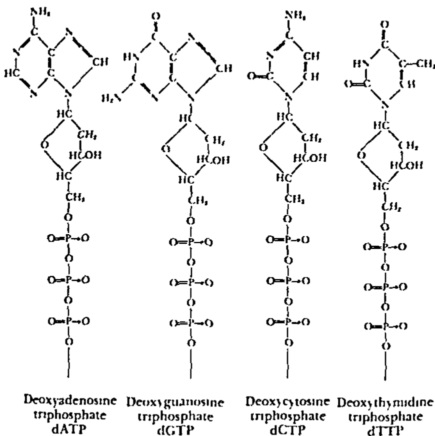


FIG 18 The structural formulas for the four compounds of base, sugar, and triphosphate which are used to synthesize DNA

which is thymidine triphosphate, illustrates one of the components of DNA, and it is assembled into the chain of DNA by breaking off the last two phosphates and joining the remaining one to the sugar of the next triphosphate to form the sequence "base—sugar—phosphate, base—sugar—phosphate . ." *which extends to millions for DNA and thousands* for RNA. To consider what happens we need to know some rates In Table 5, some rates are given.

TABLE 5. SOME RATES OF SYNTHESIS BY THE BACTERIAL CELL

DNA synthesis	3000 base pairs per second
mRNA synthesis	as high as 100,000 base pairs per second
Protein synthesis	5–50 amino acids per second per polysome
Enzyme turnover	often 1000 per second; it is very variable

Some comment on these is in order before they are used in discussion. The rate for DNA synthesis is that which is found both for bacteria when they are growing normally and optimally and for bacterial viruses. The rate for messenger RNA is hard to estimate because it is not certain whether all the DNA is synthesizing messenger at once. If it is not, then the figure given is justified.

One very simple assumption can be made regarding the process of assembly of the DNA and this is that the triphosphate must simply collide with the right place. It need not do so with the right aspect, for it is likely that rotational brownian movement will correct the faulty collision very fast, nor is it necessary for the enzymatic process to take any time after the collision. Both these factors might take some time, so the assumption made is truly an extreme representing the fastest that can possibly be done by normal chemical methods.

Now the rate of collision can be estimated in the same way that Smoluchowski did and as I described it,[9] and the very simple expression, the sole formula in this article is found.

$$\phi = 2 \pi DRC$$

where D is the rate of collision at a target spot of total radius R (this is the sum of the radius of the impinging molecule and the spot to be hit) and C is the concentration a long way away from the point of collision. Now if we substitute the known value of ϕ, we have only to know D and R to be able to estimate the concentration of the phosphates which must be present. The value of D was measured by Dr. Lehman at Penn State, for representative substances,[10] and the value of R can

be estimated with fair accuracy as it involves the radius of the triphosphates. So the concentration can be found. Attempts in our laboratory to measure the actual concentration of the triphosphates have not been too successful because they are obviously very small. Deoxyadenosine triphosphate can hardly be present in more than one twentieth of the amount needed, and the general impression is that the amount is still less. On the other hand, thymine, as the free base, is present to about the right amount. This small concentration of triphosphate leads one to wonder what is at work. Is chemistry to be abandoned or is there some way out? One way is that taken in Figure 17, where it is suggested that the machinery to make the triphosphates is very close to the polymerases, so that the concentration of the triphosphates throughout the whole cell is not that which is needed to drive the reaction at the rate observed. More work needs to be done on these actual concentrations before the argument above can be taken too literally. However, the evidence is suggestive that there is some organization at work in assembling the DNA.

If we now move on to the process of protein synthesis, this can be considered by looking at Figure 19. This is a strange, straggly-looking figure, but it still can be used to give a lot of information. The long line with the short whiskers is the messenger RNA, and the whiskers represent the bases that form the method of coding for the right protein. The first action is the attachment of a ribosome to the messenger RNA, and this promptly makes possible the attachment of a transfer RNA molecule with an amino acid on it at the end. The ribosome then moves on one "notch" and another transfer RNA comes into place with another amino acid held at the end, with its amino group and carboxyl group both held ready for the union and formation of the peptide bond. This may well be achieved by means of a peptide bond enzyme, if the evidence from Dr Schweet's laboratory[11] is right, and as we believe it, we have shown a dotted outline of this enzyme behind the

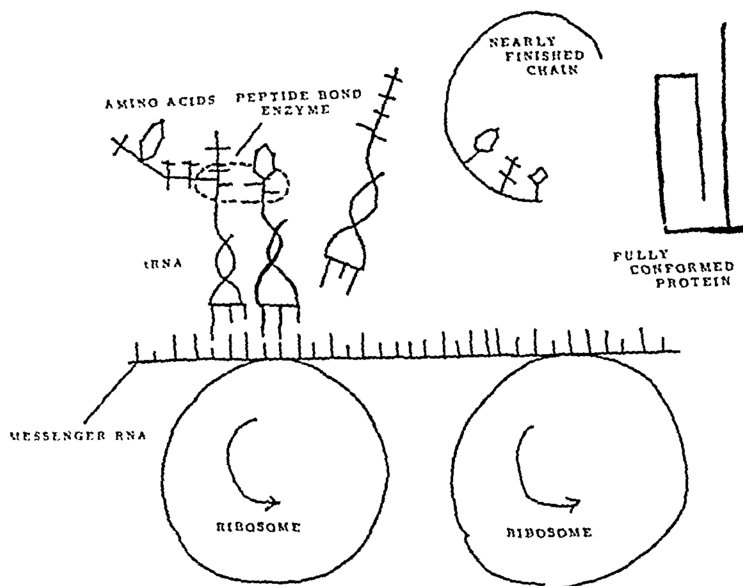


FIG. 19. A very schematic representation of the ideas involved in protein synthesis.

carboxyl and amino groups, ready to function. Another tRNA molecule is shown as being on the way. While all this has been happening, another ribosome, in fact about five more, have been working their way down the messenger RNA and each one has been manufacturing protein. One that has nearly completed its chain is shown on the right, and the final action, after the whole chain has been finished, is to *shape itself into* its final conformation, represented poetically on the far right. This is the form necessary for the enzymatic function to be effective.

Once again we have to think about rates. If the slowest probable value is taking place, then the ribosome is rolling along so that the clicks go at about the rate of a good typist typing each letter. It should be remembered that there are two typewriters going together at once: the ribosome and the peptide bond enzyme, to say nothing of the five other sets of

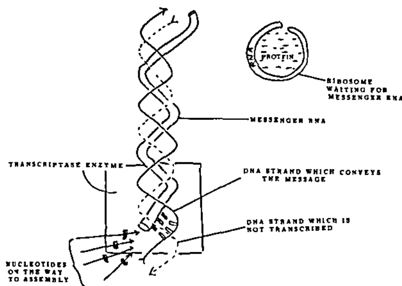


FIG. 20 The formation of messenger RNA by copying one strand of double-stranded DNA

operators farther down the messenger RNA To keep the analogy, the output is at least as good, *per polysome*, as a good, efficient, and large office If the rate is as high as it might be, then we have to think that the rolling of the ribosome and the operation of the peptide bond enzyme is like a low musical note, perhaps like the hum of the A C. As this hum progresses, think of the protein as sweeping down the messenger RNA in tune with it I find it most exciting, for it is not a single process at all but involves at least two factors operating together

Our last factory is the one that makes messenger RNA. If we accept the evidence from Dr. Berg's laboratory,[12] then both strands of DNA must be present Quite a variety of evidence suggests that only one of the strands is "copied " So in Figure 20 we have tried to show this The messenger RNA is the tube-like object formed on the continuous strand of

DNA with the dotted strand also present. The actual point of synthesis is shown, with the four bases making their way into the synthetic point. To hide abysmal ignorance, we show the shape of the enzyme as a simple rectangle. Again we need to think of speed. In this instance the enzyme is moving so fast that one can hardly dare to suggest a comparative speed. It is moving down the page, with sharp streaks of messenger following its motion, and its speed is such that it would be hard to "stop" with the best electronic flash. Somehow, a ribosome must be involved with the messenger, and we have drawn one in the vicinity to remind the reader of that fact.

A Schematic Picture of the Cell

We are now ready to see what happens when we try to put these features into the confines of a bacterial cell. The result is seen in Figure 21. It obviously calls for comment, in fact it is meaningless without, as it is intended only to focus thought. We can take the DNA first. The "old" DNA is shown as the

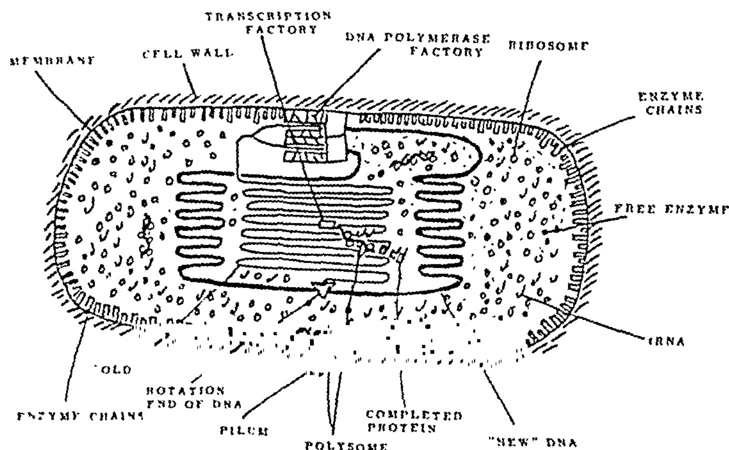


FIG. 21. The cell of *E. coli* expressed schematically, embodying the ideas developed in the previous figures

thin line, starting at the spinning mechanism of Dr. Cairns. It winds back and forth until near the top of the cell it divides and passes through the polymerase factory, which we have shown as attached to the cell membrane.[13, 14] The new DNA is shown as a thicker line and is here indicated as in two beginning daughter nuclei, getting ready to divide. It would probably be better to put these two daughter nuclei inside the old DNA so as to be pushing them outward as the cell progresses The two ends return to the spinning mechanism.

To take the formation of messenger RNA and polysomes, we have great difficulty. The process is shown as forming on the DNA, and the polysome, transfer RNA, and finished protein are indicated However, the impression so gained is too quiet and simple In reality, the whole of the outer part of the DNA must be covered with the apparatus that makes the messenger and polysomes, and almost the only way to convey the idea of this intensely active part of the cell is to suggest that the whole of the DNA is covered with a swarm of bees, actively making the products

The component parts of tRNA, free enzyme, and ribosomes are shown at each end of the cell, although there is some doubt whether there really are many of these free agents in the cell.

Comment is in order on two other features of the diagram. The first is the small whisker at the bottom, one of the pili which have been studied notably by Maccacaro[15] and Brinton[16] These are interesting because some of them are involved in bacterial conjugation and some in the process of invasion by small RNA-containing viruses, as has been shown by Edgell in Ginoza's laboratory at Penn State Each of these, at a conservative estimate, must contain 120 protein molecules, all identical, because a single mutation can remove the whole whisker Now, according to Levinthal,[17] the output of a single polysome is twenty protein molecules, after which the messenger RNA has decayed, and so more than one poly-

some must be involved in the synthesis of one of the pili. Since it is not easy to see how a very large object like a polysome can diffuse readily in the cell, it seems probable that a set of six or so messenger RNA molecules is made together very near to the place where the "pilum" is being formed. This argues a kind of deliberate action on the part of the molecular apparatus of the cell, and it is significant that this type of process must be occurring in other places. Put very strongly, the cell has a design for each of the pili, and so, very probably, for many other of its features.

The last feature that needs comment is the presence of sets of enzymes near the surface of the cell. This kind of organized set of enzymes is found in the electron transport system of mitochondria,[18] and evidence for something like it was found by Kempner and the author[19] some years ago at Yale. In looking at the radiation sensitivity of the ability of the bacterium to incorporate glucose into the macromolecular fraction of the cell, it was found that the sensitivity was far greater than that of one enzyme molecule. Now if the enzymes responsible for the whole process of assimilating glucose were all separate, then the inactivation of individual enzymes would not have much effect, for there are hundreds in the cell. On the other hand, if the enzymes are in grouped units together, then the inactivation of any one enzyme will destroy the operation of the group and the sensitivity is explained. Because the processing of metabolites should take place near the entry point of the materials into the cell, these systems of enzymes have been placed near the surface.

To continue the process of understanding the nature of the cell, two more strokes of the brush should be added to this picture. One is Figure 22, in which an attempt has been made to portray the section of the DNA right through the middle of the cell. It is roughly to scale, with the DNA being 20 Å in diameter and the separation between strands about 100 Å. The figure is very laborious to draw and it is quite easy to get

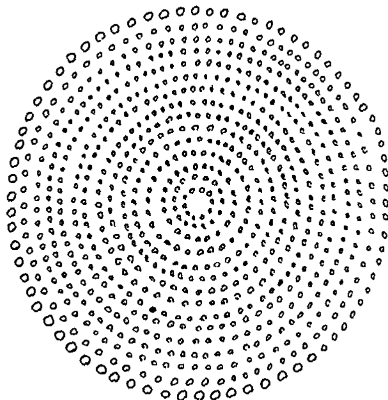


FIG 22 An attempt to show a cross section of the DNA in the cell.

lazy about it Nevertheless, it can be seen that very large numbers of strands of DNA have to be shown in this cross section, in fact about 900 of them The concentration of the DNA in this way is significant just alone, but two things need to be said about it in addition One is that, at least on the surface of the DNA where this is moving into the polymerase factory, the DNA is actually in quite rapid motion. If one were to try to represent it, the reader should draw his finger in and out just about three inches on this scale at about the rate of one in and one out every two seconds In other words, the DNA, as it moves into the replication apparatus, is moving at the rate of one micron per second Furthermore, if the transcription

process is primarily taking place in the outer layers of the DNA, then there is a very rapid peeling off of the messenger RNA all around the surface of this DNA. This remarkable activity associated with this many-convoluted kind of worm is again a very striking and curious part of the cell.

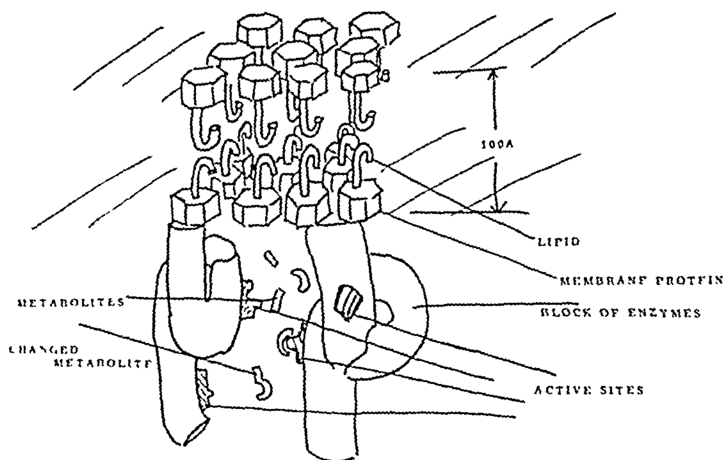


FIG. 23. The membrane and associated enzymatic structures.

The last stroke of the brush is shown in Figure 23, in which the surface of the cell has been indicated. This must, once again, be thought of as schematic and is intended to indicate the association of lipid and protein with the lipid as the little hooked candy cane-type of structures and the protein as the hexagonal larger units. This follows the suggestion of Dr. J. F. Danielli, and the separation between the layers is probably somewhere in the neighborhood of 100 Å, as indicated. Many more of these should be drawn, but the fatigue of indicating them grew great and it is better just to rely on the idea of a schematic picture. In order to understand correctly what is happening, however, one should recall the dynamics of the operation of the cell; this means that, during the time the reader has been reading the words about the membrane,

probably one more lipid molecule has appeared in the surface, and possibly one more combination of lipid and protein. One has to realize that the cell membrane is not of the nature of a balloon, which expands without changing the amount of material in it, but instead is a mechanism which actually has more and more material arranged in this layered way as time goes on. So that again, as we have continued to write this paragraph, another unit of membrane has appeared in there. It must also be recalled that, in all probability, material for this membrane must be fed from the inside of the cell outward. How to do this is of great interest, and is not solved at the present time.

Below the membrane are represented the systems of enzymes which are supposed to synthesize the necessary metabolites. These systems of enzymes are drawn in a non-geometric way to stress the fact that they are living systems, one must realize that they are in fact ten enzyme molecules or so put together. A few of the enzymatically active surfaces have been indicated, one of which, for example, will collect the curved molecule on its surface and there join to it the straight molecule to produce the sickle-shaped object farther down This is a representation of the process of the assembly of the metabolites from the materials supplied by the outside environment of the cell In thinking about this picture one should also recall that something must be going upward to the membrane as well, so that this is truly a remarkable structure.

The author cannot avoid looking at the picture and realizing that it must not only contain insight about the cell, but probably insight about himself and that it might be valuable material for a psychiatrist in explaining the workings of his mind.

It could easily be wondered whether these imaginative pictures, which portray the nature of the cell, have any reality or not. In order to establish some basis for judgment, in Fig

ure 24 is shown the picture which was drawn under similar circumstances in 1960.[20] The comparison between this picture and the one in Figure 21 is of interest, because it affords some idea of whether these pictures do represent something that is moving forward to reasonably firm conclusions about the workings of the cell, or whether they represent something that is almost random and not connected to reality at all.

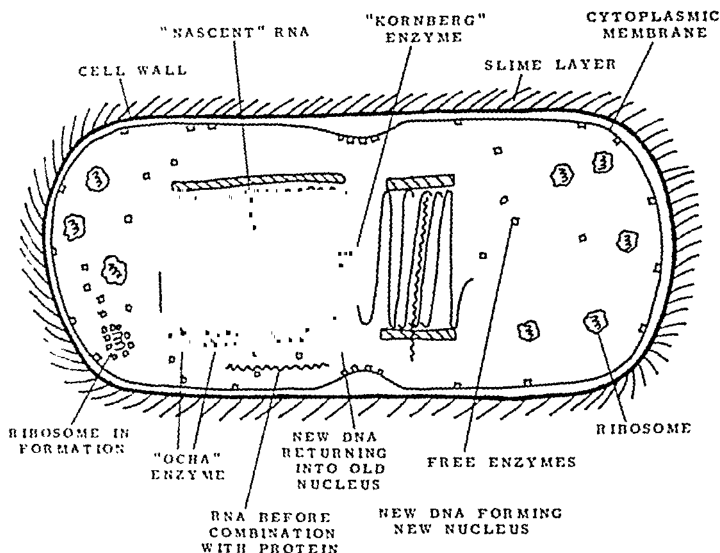


FIG. 24. An early (1960) representation of the cell for comparison with Figure 21.

In the 1960 picture, several features can be seen that are reasonably correct and several features are not so. For example, in 1960, the picture did contain one point of DNA replication, although it was not put on the cell membrane. This seems to have been before its time and quite reasonable. On the other hand, the DNA has ends, which is against the modern idea that DNA is a continuous loop. Also in 1960, the RNA is indicated as being made on the DNA, which is

reasonable in the view of today's thinking, but in 1960 there was no concept of messenger RNA, and in fact our thoughts in these "old" days were that the RNA was made on the ribosomes and the ribosomes proceeded to make the protein without any assistance from messenger. This is not held to be true today. One can therefore see that these pictures do seem to represent a steady advance, and it does not seem to the author to be unreasonable to make the claim that if, in 1970, a group of scientists resident in Japan, Russia, France, the United States, and Sweden were brought together by a UNESCO conference to compare their pictures of the same cell, that these pictures would look remarkably alike and that they would really bear some relationship to the truth.

The author would therefore like to end this section on the optimistic note that the inferential method, or the method of accurate imagination, does have the ability to tell us something about the structure of the cell and that this schematic drawing which we have just produced and the discussion we have used is really something that we can consider firmly in our idea of how to go about attempting to synthesize such a cell

Ways of Synthesizing a Cell

It is now our task to try to find ways in which this complicated structure could be synthesized We can consider one very quickly and dismiss it. If it is necessary to assemble the cell we have just described as we would assemble a watch, that is to say to make separately every individual component, to put them together and then, so to speak, pat it on the back and it will start up, then I do not think we are ever going to do it. I do not believe we are going to be able to make the DNA in one long strand, or thread it through the polymerase apparatus, pull it out on the far side, loop it around, put the transcription mechanisms on the outside, encase it in its beau-

ure 24 is shown the picture which was drawn under similar circumstances in 1960.[20] The comparison between this picture and the one in Figure 21 is of interest, because it affords some idea of whether these pictures do represent something that is moving forward to reasonably firm conclusions about the workings of the cell, or whether they represent something that is almost random and not connected to reality at all.

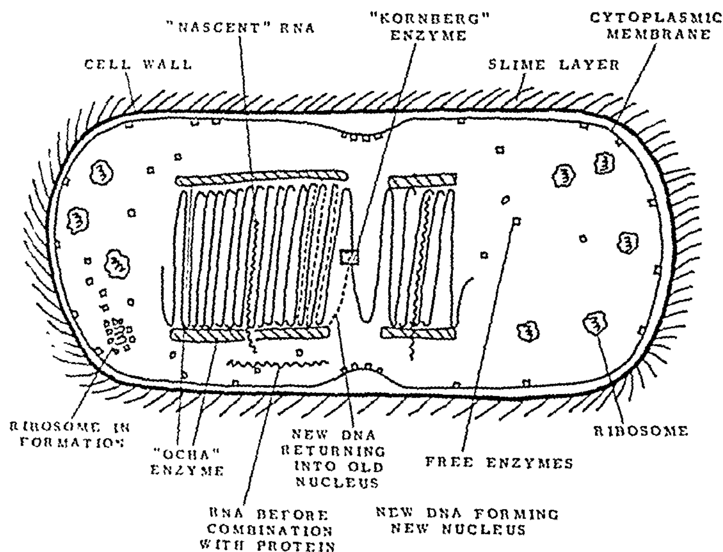


FIG. 24. An early (1960) representation of the cell for comparison with Figure 21.

In the 1960 picture, several features can be seen that are reasonably correct and several features are not so. For example, in 1960, the picture did contain one point of DNA replication, although it was not put on the cell membrane. This seems to have been before its time and quite reasonable. On the other hand, the DNA has ends, which is against the modern idea that DNA is a continuous loop. Also in 1960, the RNA is indicated as being made on the DNA, which is

reasonable in the view of today's thinking, but in 1960 there was no concept of messenger RNA, and in fact our thoughts in these "old" days were that the RNA was made on the ribosomes and the ribosomes proceeded to make the protein without any assistance from messenger. This is not held to be true today. One can therefore see that these pictures do seem to represent a steady advance, and it does not seem to the author to be unreasonable to make the claim that if, in 1970, a group of scientists resident in Japan, Russia, France, *the United States, and Sweden were brought together by a* UNESCO conference to compare their pictures of the same cell, that these pictures would look remarkably alike and that they would really bear some relationship to the truth.

The author would therefore like to end this section on the optimistic note that the inferential method, or the method of accurate imagination, does have the ability to tell us something about the structure of the cell and that this schematic drawing which we have just produced and the discussion we have used is really something that we can consider firmly in our idea of how to go about attempting to synthesize such a cell

Ways of Synthesizing a Cell

It is now our task to try to find ways in which this complicated structure could be synthesized We can consider one very quickly and dismiss it If it is necessary to assemble the cell we have just described as we would assemble a watch, that is to say to make separately every individual component, to put them together and then, so to speak, pat it on the back and it will start up, then I do not think we are ever going to do it. I do not believe we are going to be able to make the DNA in one long strand, or thread it through the polymerase apparatus, pull it out on the far side, loop it around, put the transcription mechanisms on the outside, encase it in its beau-

tiful membrane with the enzyme structures close to it, and then see it start. I believe we may definitely dismiss this right at the outset.

However, to take another analogy, there is a way in which we might think about it. Perhaps we could make the cell as we make a crystal. If we wanted to make a crystal of alum, we would make a concentrated alum solution, hang a small seed crystal inside and go away, and the following day we would find a nice crystal and say we had made it. In a certain sense, we had not made it, and yet we would think we had. What would have happened is that the crystal "wanted" to form, or in scientific terms there were forces that would develop the order present in the crystal, and these forces took over and made it under the right circumstances. The question arises whether there are forces that would take over and make the living cell if we started it in the same way that we could start a crystal forming. This is a tremendously important question which has roots in both philosophy and in the extent to which our knowledge of the laws of nature is still limited. Before discussing these philosophical points, it might be worth while to look at something slightly less formidable than the concept of actually making a start on a living cell and having it finish itself. What is much easier to contemplate is the idea that we might be able to make a primitive form of a living cell and then hurry its evolution. If we are going to make a living cell, and if the analysis here given is of any use, then the method which suggests itself as the best is the one just described. Find out what a primitive cell is like, make that, and then hurry up the evolution by deliberately producing the things we know will one day lead to the proper present-day bacterial cell.

There is some real hope that we might be able to find out something about the primitive cell in the recent work of Hoyer, McCarthy, and Bolton.[21] Their work contains a suggestion that the DNA in cells has been essentially overpre-

served—that the necessity possessed by the cell to keep its DNA intact may actually have caused so great a restriction on its degradation or removal that DNA has accumulated in many cells in such a way that we can look at it even though it is today not very functional. If this should prove to be true, and it is certainly an attractive hypothesis, then we will find stretches of DNA in present-day bacterial cells which represent much earlier stages in its existence. We might be able to infer from these what would be the most primitive we could "get away with" and from the inference we might be able to start the construction of the DNA necessary to operate as an extremely primitive cell. It could be that, in fact, cells in their very early stages were not built to use the "holy trinity" of DNA, tRNA, and protein. We would have to make our conclusions with regard to that and start trying to assemble the very simplest possible kind of thing, and then by knowing the way in which mutations have occurred, by alteration and extension of the DNA, so add the necessary parts as to bring the cell into being much faster than the thousands and millions of years it must have taken. This is the method which is suggested as the way in which the cell might be made.

It is now necessary to comment on what would be involved. It is legitimate to try to use intuition in making suggestions and predictions. The author well remembers an occasion, when he was an impressionable graduate student, when Lord Rutherford outlined his theory of the atomic nucleus.[31] This was in 1927, five years prior to the discovery of the neutron. In his theoretical account, Lord Rutherford employed already outdated forms of quantum mechanics; he used the modification of the Bohr-Sommerfeld theory with half-quantum numbers, and he used a great many clear infractions of the best theory of the day. It did not fare well, and I remember the polite but insistent grilling to which this g[entle]man, at that time essentially twice a Nobel laureate, presi[dent] of the Royal Society, recipient of the Order of Merit, .

perhaps the greatest figure in science of his day, was subjected. I remember Lord Rutherford's face grew redder and redder until finally he brought his hand down on the table and he said "Gentlemen, I feel it in my bones that there are neutrons in the nucleus." It is hard to say whether his theory was a good theory or a bad theory in the light of the fact that he was right on this major premise. What we are asking ourselves is what do we "feel in our bones" with regard to the living cell and the problem imposed with regard to its synthesis. We have already discounted the idea that we could put it together like a watch, and we have suggested the best way would be to think about an early cell, to make it, and then to accelerate its evolution. The question is, does the cell automatically contain something which has an orderly and regular behavior, even though it has complex components, even at the most primitive stage at which we could possibly recognize a cell? This is a very crucial question and one that is at the heart of the entire situation we are discussing. If it is indeed possible that the semichaotic behavior of chemical systems can lead progressively by ramification upon ramification to something as complex, as ordered and, one might almost say, as "thinking" as the cell we have just described, then we should indeed press resolutely forward to show that this is true. And if it is, then there is no question that the greatest triumph for our knowledge of our laws of nature will be in this successful synthesis by what we know at present of a living cell. On the other hand, if it should prove that all our attempts to do this fail, and yet we are faced with the reality of living systems in abundance growing and developing and behaving in the way they do, then we will be forced to contemplate something new and something different about life which is not present in our scheme of nature as it is known today. The question arises whether this type of new thing, which is inherently shown at its maximum in living things, though one would hope that it penetrated the whole of nature, is some-

thing we must look for and first find before we can make a living cell de novo. The sharp philosophical question is not resolved today nor does it seem simple to suggest the way in which it will be resolved. There is only one good way to go about finding it, and that is by designing experiments resolutely in terms of the concept that the present laws of physics, and of physical chemistry are indeed adequate to explain all living things. Using this idea, we then push resolutely with this explanation and conduct experiments in those areas where the explanation appears to be marginal Thus, it would seem to be wise, if one is really interested in this kind of question, to look not only at the structure of cells but at the way the structure is changing with time In other words, it is the rates of synthesis perhaps, and not the actual accuracy of synthesis, which may pose the challenge. Put more succinctly, since living things do contain the fourth dimension, we must study all four dimensions if we study living things with great care.

Whether these studies and this attitude will result in a completion of the project mentioned at the beginning of this paper, namely the logical assembly in our minds, step by step, of the way to put together a living cell, or whether we will find steps that are missing that cannot be completed, is still open to question It is because this question is open that the work is so exciting and so demanding and that one does feel this is indeed the great frontier in science today.

In one concluding remark, I would like to return to the analogy of the scientific revolution that occurred early in this century. It is true that the essential understanding of inanimate nature was obtained within forty years of the discoveries made at the turn of the century However, in making these discoveries, even though great adaptations had to be made early, as in the case of the Bohr theory, the final confrontation of good accurate theory with the problem of two electrons in an atom led to one of the greatest surprises in physical

science of all time, namely the idea that we must substitute an explanation in terms of probability for an explanation in terms of continuous certainty. This substitution did not diminish our skill and understanding: rather it increased it. We are now able to make confident predictions about atomic behavior, about chemical structures which are based on this theory. Nevertheless, it must be conceded that the theory as it came out produced not only surprise but, it would be fair to say, shock.

The question that is fairly before us, and posed by the problem of the artificial synthesis of a living cell is the question whether life and its understanding will also bring with it to the scientific community, now as a whole, the same degree of shock that was produced then.

REFERENCES

A great many competent research workers have created the field of molecular biology which is partly described here. To do correct justice to the development of each topic is beyond the author and not really required in covering the subject of cell structure. To supplement the numbered references below, two general references are suggested here, from which a more complete appraisal of the subject can be obtained.

J. D. WATSON, *Science, 140*, 17 (1963).

This Nobel Prize address covers many topics with care and clarity and is very useful reading.

Cold Spring Harbor Symposia, on Quantitative Biology, Vol. 28 (1963). (1963).

This most important volume is the closest one can come to a complete and authoritative account of the molecular biology referred to in this article.

1. G. V. LEVIN and N. H. HOROWITZ, Report for Hazleton Laboratories Inc., Contract NASr-10, Tables 11-15 (1964).
2. A. K. KLEINSCHMIDT, D. LANG, D. JACHERTS, and R. K. LAHR, Darstellung und Langenmessungen des gesamten Desoxribonucleinsaure-Inhaltes von T_2-Bakteriophagen, *Biochim. Biophys. acta, 16*, 857 (1962).

3 J CAIRNS, The Bacterial Chromosome and Its Manner of Replication as Seen by Autoradiography, *J Mol Biol*, *6*, 208 (1963)

4. SCHRODINGER, *What is Life?* Cambridge Univ. Press (1962).

5 R W HOLLEY, J APGAR, G A EVERETTE, J. T. MADISON, M. MARQUISEE, S H. MERRILL, J. R PENSWICK, and A ZAMIR, Structure of a Ribonucleic Acid, *Science*, *147*, 1462 (1965)

6 H QUASTLER, *Information Theory in Biology*, Univ of Illinois Press (1963)

7 H YOSHIKAWA and N SUEOKA, Sequential Replication of *Bacillus subtilis* Chromosome, I Comparison of Marker Frequencies in Exponential and Stationary Growth Phases, *Proc Nat Acad Sci U S*, *49*, 559 (1963)

8 K G LARK, T REPKO, and E J HOFFMAN, The Effect of Amino Acid Deprivation on Subsequent Deoxyribonucleic Acid Replication *Biochem Biophys Acta*, *76*, 9 (1963)

9 E C POLLARD, Collision Kinetics Applied to Phage Synthesis, Messenger RNA, and Glucose Metabolism, *J Theoret Biol*, *4*, 98 (1963).

10 R C LEHMAN and E C POLLARD, Diffusion Rates in Disrupted Bacterial Cells *Biophys J*, *5*, 109 (1965)

11 [illegible]

12 M CHAMBERLIN and P BERG, Studies on DNA-Directed RNA Polymerase, Formation of DNA–RNA Complexes with Single-Stranded ϕX 174 DNA as Template ibid, p 67

13 A T GANESAN and J LEDERBERG, Membrane-Bound Fraction of Bacterial DNA, *Biophys Soc Abstr*, 162 (1965)

14 D W SMITH and P C HANAWALT, State of Aggregation of the Growing Point in the Bacterial Chromosome, *Biophys Soc Abstr*, 162 (1965)

15 G A MACCACARO and W HAYES, The Genetics of Fimbriation in *Escherichia coli*, *Genet Res*, *2*, 406 (1961)

16 C C BRINTON, JR, Non-flagellar Appendages of Bacteria, *Nature*, *183*, 782 (1959)

17 C LEVINTHAL, A KEYNAN, and A HIGA, Messenger RNA Turnover and Protein Synthesis in *B subtilis* Inhibited by Actinomycin D, *Proc Nat Acad Sci U S*, *48*, 1631 (1962)

18 A L LEHNINGER, *The Mitochondrion*, Benjamin, New York (1964)

19. E. S. KEMPNER and E. C. POLLARD, Radiation Inhibition of Amino Acid Uptake by *Escherichia coli*, *Biophys. J.*, *1*, 265 (1961).
20. E. C. POLLARD, Theoretical Aspects of the effect of Ionizing Radiation on the Bacterial Cell, *Am. Naturalist*, *94*, 71 (1960).
21. B. H. HOYER, B. J. MCCARTHY, and E. T. BOLTON, Complementary RNA in Nucleus and Cytoplasm of Mouse Liver Cells, *Science*, *140*, 1408 (1963).
22. E. RUTHERFORD, Structure of the Radioactive Atom and Origin of the Alpha Rays, *Phil. Mag.*, *4*, 580 (1927).

7. THE STRENGTH AND WEAKNESS OF BRITTLE MATERIALS

— By C. J. Phillips
Rutgers, The State University

The behavior of brittle materials is much like the mythological box of Pandora. Both are full of surprises, contradictions, hopes, and frustrations.

> There is a strange fascination about a mythological character that has retained its vitality up to our own day and, in the course of time, has lent its name to English queens as well as to French policemen, to the philosopher's stone as well as to a gang of fifteen-year-old murderers in Philadelphia.[1]

In one version of the story Zeus gave Pandora a box, most beautiful, but which he forbade her to open. As most women might do she opened it anyway, thus letting out all the evils which since have attended man. Only one thing, hope, remained inside.

Let us now look at the behavior of a typical brittle material, glass. We know how to treat specially and to test fairly sizable silica glass fibers and attain tensile strengths exceeding 2,000,000 psi, a fantastic value. It is not too difficult to etch with acid much larger glass rods and break them at tensile stresses over 500,000 psi, again a very high strength. New

chemical tempering methods, applied to quite massive glass objects, result in moduli of rupture greater than 100.000 psi, still quite a respectable figure. But now the paradoxical nature of the situation appears. Some of the treatments needed to attain these high strengths are completely impractical from any present commercial point of view. All the articles so treated are degraded in strength, usually to a small fraction of the original value, by the kind of handling that is necessary to make them into objects of utilitarian value. We have no practical way to prevent the scuffs, scratches, abrasions, and chemical attack that come from making and using glass products. Surface coatings—lacquers, paints, silicones, and plastics—are expedients which may temporarily alleviate the problem, but they do not solve it. Thus, we find most glass products breaking at stresses of 2,000 to 20,000 psi—seldom over 50,000 psi—two or more whole orders of magnitude lower than we believe their theoretical strength should be. When we leave the lid closed on Pandora's box, all is well. As soon as we open it, we destroy the very things we have created.

A Proper Frame of Reference

In recent years excellent reviews have appeared which cover the behavior of brittle materials, and some of these are not restricted to glass. Overall, they include the publications of Jones,[2] Orowan,[3] Shand,[4] Walton,[5] Averbach et al.,[6] Charles and Fisher,[7] Charles,[8] Hillig,[9] and Ernsberger.[10] Consequently, in this paper, the only references cited will be those considered of prime importance, or those that are necessary to the continuity of the discussion.

A very proper question at this point would be: Why another review now? There are really five reasons. The most obvious is that it updates the information by at least another year or two. The second is that it presents the data to a circle of readers with broader and more diversified interests than

those to which the publications listed above might appeal. The next reason is simply that it is written by another person, with a different point of view and with whatever plus and minus factors this may include. The fourth reason is somewhat more subtle. It has to do with a proper frame of reference. Ernsberger began his review with this statement:

> It is probably safe to say that more tangible progress has been made in our understanding of the strength of glasses within the last five years than was made in the entire preceding period of thirty-five years dating back to the initial formulation of the Griffith crack hypothesis.

In certain areas, and these include Ernsberger's own beautiful work, this is indeed true In other areas, it probably is not. For example, Schurkow[11, 12] had already broken fused silica fibers in vacuum at stresses ranging from 1,700,000 to 2,300,000 psi, stresses as high as we have obtained to date In the period 1931-36, Greene and Phillips, working at Corning Glass Works, had already produced sizable acid-etched commercial glass rods which broke at stresses exceeding 400,000 psi One ¼-in rod withstood a stress of 410,000 psi for one hour before it broke. With three-point loading on an 8-in span, the deflection at the center was approximately 1.75 in Another rod, ⅜ in. in diameter, withstood a stress of 300,000 psi for several days

It is hard to believe these are pieces of "brittle" material This part of the story is brought up to date, 28 years later, by Figures 30 and 31. Thirty years ago the effects of surface abrasion, static fatigue, area under load, and testing in vacuum were all appreciated, if not thoroughly understood It did not seem expedient at the time, for good and sufficient business reasons, to publish this information It is brought up now, not to establish priority or to be contentious As the Italian proverb says, "The graveyards are full of alter-

wisdom." But it does point up the fact that, in certain directions, not as much recent progress has been made as one might desire. The last reason for the present review is that it presents certain new data in a form that may be interesting and stimulating to other workers in this field.

Brittleness and Ductility

The word "brittle" appears in the title of this paper but we have not as yet defined it. Nor is it easy to do so. We might say that any material is brittle which obeys Hooke's law right to the breaking point without any permanent plastic deformation whatsoever. This means that as soon as the stress is applied to such a material the atoms or ions shift to new off-center locations, but no bonds are broken until the ultimate strength is exceeded. Removal of any lower stress permits the atoms or ions to return immediately to their original low-energy positions without any permanent change in structure. If, on the other hand, at stresses below the ultimate, bonds are broken and the atoms or ions move to new locations, plastic deformation has occurred and some or all of it will remain after the stresses are removed. Such a material is called "ductile." This sharp distinction between the two types of behavior is gradually disappearing because it is now realized that, in "real" (as distinguished from "ideal") materials, there is considerable overlapping. For example, given the appropriate geometrical constraint, rate, and type of loading, and ambient temperature, such an apparently ductile material as a copper alloy may be made to exhibit a brittle type of fracture, as in a notched bar impact test. The reverse is also true: materials normally considered brittle may sometimes exhibit ductility prior to fracture. Nevertheless, the differences are far greater than the similarities, and there will seldom be confusion between the two. In fact, Dr. L. C. Tyte,[13] comments on the quite striking agreement in the

behavior of six widely different brittle materials: brick, coal, concrete, gypsum, rocks of different kinds, and glass. This behavior is normally very different from that displayed by ductile substances.

Ceramics and Choice of Glass for Detailed Study

Ceramics may be briefly defined as products made from nonmetallic, inorganic materials, generally involving high temperatures during their manufacture or use. They run the gamut from single crystals at one end, through polycrystalline materials which may or may not have a substantial glassy phase, and finish at the other end with glasses which are usually considered completely noncrystalline, i.e amorphous. With trivial exceptions, all these ceramic materials are brittle in the sense described above As such, they have received much less study and attention than the materials called ductile, notably the metals. This may be because they are manufactured in much smaller quantities, they are often extremely complex in composition, and they are sometimes quite inhomogeneous A few years ago, Bradstreet[14] semifacetiously described the manufacture of ceramics as a process in which

> minerals of inconstant composition and doubtful purity are exposed to immeasurable heat long enough to carry unknown reactions partly to completion, forming the heterogeneous non-stoichiometric materials known as ceramics

Although much progress has been made in recent years in controlling some of the variables Bradstreet mentions, we would still be wise in this review to seek out first a ceramic in which some of these variables are already minimized Such a material is glass. It is also a classic example of a brittle material par excellence. Like most brittle solids, glass usually

breaks at stresses less than 1 per cent of the theoretical ultimate strength. But, in glasses, we feel quite sure that the imperfections which reduce strength are confined to the *surface*. In crystals, on the other hand, imperfections may include dislocations, interstitials, vacancies, stacking faults, and impurity atoms. For polycrystalline materials, we must add internal cavities and grain boundaries. We can simplify our task considerably if, initially at least, we confine our study to glasses and, in most cases, to silicate glasses. It is worthy of note that much more work has been done on the strength of glass than has been done on any other ceramic. It was also the material chosen by Griffith[15] for his pioneering work on strength. His theory is still viable almost fifty years later and is applied to rupture in brittle solids of all kinds.

The Structure of Glass

An abbreviated but reasonably precise definition of glass is that it is "an inorganic product of fusion which has been cooled to a rigid condition without crystallizing." At high temperatures, glasses are true liquids, but liquids of great viscosity. At 2800°F a typical glass might have the same viscosity as honey or molasses at room temperature. This viscosity increases continuously and very rapidly as the temperature drops. It is this tremendous increase in viscosity—by roughly 100 billion times as the temperature falls from 2800°F to 800°F—combined with rapid cooling as the glass is shaped, that serves as the principal impediment to crystallization. At low temperatures glasses appear to be hard, rigid solids. As a matter of fact, they are actually in a state intermediate between a liquid on one hand, where no permanent atomic neighbors exist, and a crystalline lattice, on the other hand, where all neighbors are fixed in perfect orderliness. In this amorphous or vitreous form of matter permanent neighbors do exist, but there is also a certain disorder characteristic of the liquid state from which the glass was born.

The Random Network Theory

In 1912, X-ray diffraction in crystals was first observed. W. H. and W. L. Bragg, father and son, immediately began an intensive X-ray investigation into the nature of solids for which, in 1915, they were jointly awarded the Nobel Prize. Before their work ended, they had shown that in hundreds of silicates, despite their complex chemical composition, there was a beautiful simplicity and orderliness in basic structure. Invariably, there were four oxygen ions arrayed around a central silicon, thus forming SiO_4 tetrahedra. The corners between the tetrahedra could be shared in a variety of ways, and the overall structure could be quite complicated This need not concern us here. What is important is that, in all crystalline forms of SiO_2 itself, the basic units of structure are these same tetrahedra with the Si–O distance varying between 1.5 and 1 7 Å, depending on the particular type of crystal Each oxygen atom, in turn, is shared by two silicon atoms so that the structure may be thought of as built up by pairs of tetrahedra which share the corners No two tetrahedra meet at more than one corner, but all corners of all tetrahedra are actually shared. A very spacious three-dimensional structure may thus be built up in several ways, and the Si–O–Si lines may or may not be straight Such a structure can be indicated by the top view in Figure 25 The fourth linkage to each Si is omitted so that the structure can be represented here in only two dimensions The X-ray diffraction pattern for such a crystal shows sharp diffraction lines The X-ray pattern of fused silica, on the other hand, shows diffuse bands rather than sharp lines Following the pioneering work of Zachariasen[16] and of Warren[17] and their coworkers, the random network theory interprets these diffuse patterns by postulating that SiO^4 tetrahedra are still the basic building blocks in silica glasses but that there is also a considerable degree of disorder, as shown in the lower view in

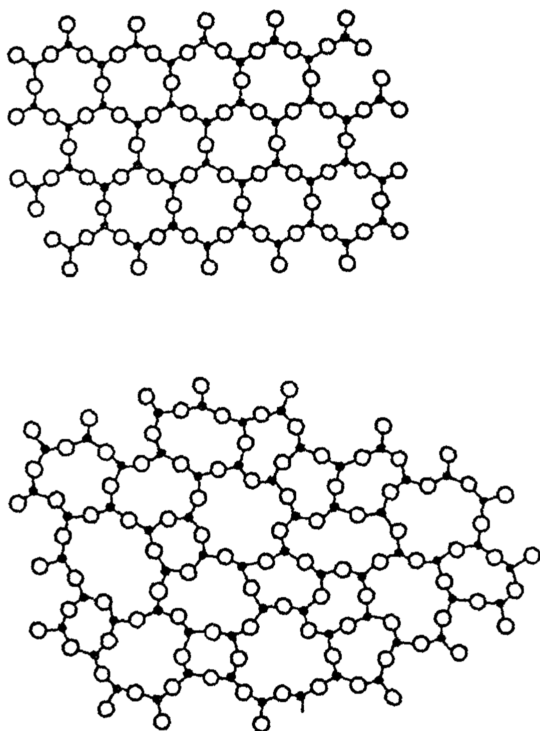


Fig. 25. Top: Schematic two-dimensional representation of a crystalline form of SiO_2; bottom: corresponding glassy form of SiO_2.

Figure 25. Fourier analysis of the spectra of glassy silica does indeed show an average Si–O distance of 1.62 Å, in good agreement with data for silica crystals. This is only a statistical order, however, and all quantitative description of the structure stops at about 7 Å from each Si as an origin. As other atoms or ions, such as Na or Ca, are added to the basic silica structure it is believed they occupy positions such as those shown in Figure 26. The nonbridging oxygen atoms thus created carry a negative charge, and the metallic ions are

held in place by ionic forces. It would appear that the introduction of any network modifier of this kind must weaken the structure by breaking some of the primary Si–O network linkages. Evidence will be given later in this paper that this may not always be true.

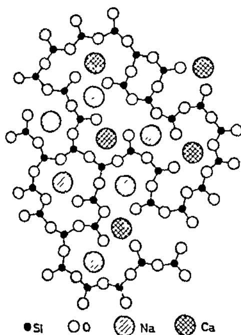


FIG 26 Two-dimensional representation of a soda-lime–silica glass (after Biscoe)

Other Theories of Glass Structure

In recent years, structural heterogeneities have been discovered in glass which cannot always be satisfactorily explained by the random network theory. These discoveries are the result of new or improved techniques in X-ray analysis, electron microscopy, infrared spectrometry, and the use of other tools not available thirty years ago We will not dwell

upon them in detail here because the literature is already quite extensive. In one review by Prod'homme,[18] thirty-eight relevant references are cited. Jellyman[19] has excellently summarized the overall situation as it exists today. He says:

> Zachariasen's and Warren's concepts of the glassy state are still valuable as working hypotheses, but most investigators on the subject today have reservations as to whether there is complete randomness of structure down to the atomic scale. It may well be that glasses contain more-highly-organized and less-highly-organized domains of dimensions up to a few hundred Ångstroms and that variations in chemical composition can occur on this scale.

It is ironic that, in many ways, we know more about the structure of man-made materials of recent origin than we do of glass, which has been made for thousands of years.

Estimates of Theoretical Strength

Brittleness, fortunately, is not synonymous with weakness. All theories agree that silicate glasses should be very strong. The strength of their chemical bonds calls for it. The various theories do not agree very well with one another, quantitatively, but none predicts an ultimate or theoretical strength less than 1,500,000 psi, and some predict strengths of 5,000,000 psi.

The mathematical equations developed in estimating maximum strength will depend upon the type of force–separation function that is employed. The simplest, and crudest, is simply to estimate that a strain of 0.1 to 0.2 would be required before a plane of atoms would be completely removed from the attractive force of its neighbors. Rupture would then be expected at a maximum stress of $(0.1-0.2)E$, where E is

Young's modulus. A second method should give an improved estimate As Gilman[20] remarks,

> Common sense tells us that the net stress between the atoms in a crystal must be zero in the absence of applied stresses, must rise to some maximum value as the atoms are pulled apart, and then must fall to zero again as the crystal breaks into two or more pieces.

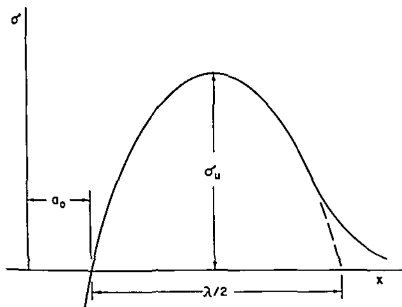


FIG. 27 Simple diatomic force-separation diagram

The same reasoning certainly can be applied to glasses, particularly if they are simple SiO_2 glasses The force–separation diagram might then look like Figure 27, where a_0 is the equilibrium separation between Si and O, σ_u is the ultimate stress required to pull them apart, and $\lambda/2$ is the effective "range" of the interatomic forces The initial part of the curve can be approximately represented by

$$\sigma = \sigma_u \sin \frac{\pi x}{\lambda/2} = \sigma_u \sin \frac{2\pi x}{\lambda}$$

For small increases in the distance between the atoms

$$E = \frac{\text{stress}}{\text{strain}} = \frac{\sigma}{x/a_0} \quad (1)$$

so

$$\frac{d\sigma}{dx} = \frac{E}{a_0} \quad (2)$$

In this same part of the curve

$$\frac{d\sigma}{dx} = \frac{2\pi\sigma_u}{\lambda} \quad (3)$$

Equating (2) and (3), and solving for σ_u,

$$\sigma_u = \frac{E\lambda}{2\pi a_0} \quad (4)$$

If the Si–O linkages supply the main cohesiveness (and they may *not* in complex glasses), we would expect λ/a_0 to remain relatively constant. If we say that $\lambda/2 = a_0$, which is not unreasonable, then $\sigma_u = E/\pi$. For fused silica $E = 730$ kilobars or about 10,600,000 psi and σ_u would be about 3,400,000 psi. As we use more sophisticated force–separation assumptions we get other equations, but the value just calculated is startlingly close to the estimate of 3,400,000 psi by Hillig[9] and of 3,500,000 psi from the very complicated model of Naray-Szabo and Ladik.[21]

Returning to the simple model of Figure 27 we can strike an energy balance between the energy $2S$ of the two new surfaces produced, and the work done per unit area during fracture, which is

$$\int_0^{\lambda/2} \sigma_u \sin(2\pi x/\lambda)\,dx = \frac{\sigma_u\lambda}{\pi} = 2S \quad (5)$$

But, from equation (4), $\lambda = 2\pi\sigma_u a_0/E$, so that when we substitute this in equation (5) and solve for σ_u we get

$$\sigma_u = \left(\frac{SE}{a_0}\right)^{1/2} \qquad (6)$$

Charles[8] summarizes the methods of estimating the surface energy S for fused silica and arrives at a value of 1,750 ergs/cm² If a_0 is taken equal to the equilibrium Si–O separation of about 1 6 Å, we get a theoretical strength of 2.8×10^{11} dynes/cm² or about 4,000,000 psi, a value not very different from those given above. It would appear that we must somehow explain, not the strength of glass, but the weakness of glass

Inglis and Stress Concentration

The solution by Inglis[22] of the notch problem was the first step toward relating observed failure stresses to ultimate strengths He showed that irregularities or flaws could act as stress concentrators so that stresses at the flaw tip could be very much greater than the locally applied stress, and might actually exceed the ultimate strength of the material and permit flaw propagation and fracture

Inglis first worked with the two-dimensional situation where the flaw could be considered as an elliptical hole in a plate under tension as in the top portion of Figure 28 He found that the stress was greatest at the ends of the ellipse where the radius of curvature is smallest This stress σ_m was related to the applied stress σ_a by the equation

$$\frac{\sigma_m}{\sigma_a} = 1 + \frac{2a}{b} \qquad (7)$$

where a and b are half the major and half the minor axes of the ellipse He further showed that, if the crack was a very elliptical two-dimensional notch, as in the lower part of Figure 28, equation (7) would still apply He finally showed that if the flaw was irregular in shape, but narrow, with [illegible]

the tip elliptic in shape, the stress concentration would be approximately

$$\frac{\sigma_m}{\sigma_a} = 2\left(\frac{a}{\rho}\right)^{1/2} \tag{8}$$

where a is now the crack *depth* and ρ is the radius of curvature at the tip. It is easy to see that if $a = 10^{-3}$ in. and $\rho = 10^{-7}$ in., the stress concentration factor is 200. If the applied stress σ_a (which is what we measure) is 20,000 psi, the actual stress σ_m at the crack tip is 4,000,000 psi and this may be sufficient to cause rupture.

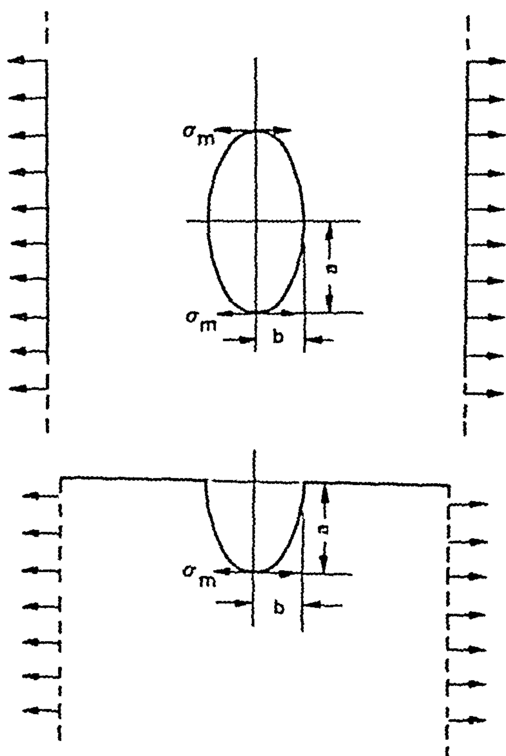


Fig. 28. Stress-concentration at elliptical holes and notches (after Inglis).

crack tip radii such that the stress values produced are equal to those actually existing where the bonds are undergoing separation. Elliot[23] and Gurney[24] have studied the situation at the atomic level. They have suggested some correction factors, but in view of the no doubt greatly oversimplified assumptions made by Griffith and all other workers in this field, it cannot be said that Elliot and Gurney or other investigators have basically changed Griffith's original concept. The sole exceptions to this are the "flaw-genesis" theory of Poncelet[25] and Marsh's theory[26, 27] of plastic flow at the crack tip, which are discussed later in this review.

It will be noted that the Inglis equation makes σ_n proportional to $\sqrt{1/\rho}$. In any real crack, the question is how to define ρ. It would appear that it could assume any of a very wide range of values, depending on kind and degree of mechanical surface damage, chemical attack, etching, etc. On the other hand, the Griffith equation does not explicitly involve ρ at all. Orowan[28] believes that the Inglis and Griffith criteria are essentially identical, simply expressing the same facts in two different forms. Hillig[9] does not agree. He believes equation (9) is *necessary* for failure but is not sufficient, whereas he considers equation (8) *both necessary* and *sufficient*. It is interesting that if we combine the two we get

$$\sigma_m = \left(\frac{8ES}{\pi\rho}\right)^{1/2} \tag{10}$$

and now the crack depth has completely disappeared! Although the preceding exclamation point indicates surprise, Hillig is *not* surprised that the crack depth disappears. He feels that equation (9) is really an inequality:

$$\sigma_n \geqslant \left(\frac{2ES}{\pi a}\right)^{1/2}$$

As a result, equation (10) has meaning only when ρ has the dimensions of an atomically sharp crack. If we do let $\rho = a_0$

we see that this is $\sqrt{8/\pi}$ or 1.6 times larger than the σ_u from equation (6).

Working with flaws very much smaller than the artificial cracks used by Griffith, numbers of people, in a number of different ways, have apparently confirmed the inverse square root relationship of equation (9) between stress and crack depth. These include Tillett,[29] Charles,[30] Levengood,[31] Mould and Southwick,[32] Shand,[33] – all with glass – and Berry[34] with glassy organic polymers Later in this paper a new relationship will be shown which will involve both E and the crack depth.

The Observed Strength of Silicate Glasses

Let us now review the basic mechanical behavior of glass, as we encounter it in everyday life, and see if we can relate it to the theories we have just discussed

The condition of the surface First, it should be noted that fracture in glass is *always* initiated by a tensile stress Even under torsion or compression the initial fracture is at right angles to the maximum tensile component. We can be quite certain of this because glass fracture surfaces, such as that shown in Figure 29, show characteristic ridges, ripples, and other fine structure which can often be interpreted to show the exact origin of the break, the path it followed, and the velocity with which it proceeded These patterns are the fingerprints of the fracture process. Except when there are gross inhomogeneities in the interior, the break always starts from the *surface* and always in an area under tension. The condition of the surface is therefore of prime importance. *How important* is shown in Table 6 These are results obtained at Corning Glass Works in the period 1931-36 but they have been repeatedly confirmed by many investigators since that time. All values are for rods of a single composition which finally broke, in a three-point bending test, after one hour



FIG. 29. Fracture surface of a glass rod (after Poncelet).

under the stress shown. Figure 30 shows a rod loaded to about 125,000 psi, which was still unbroken in 1962, twenty six years after it was first stressed. Figure 31 shows that when the stress was released, the rod returned to its original straightness within the limits of measurement. There was absolutely no evidence of permanent deformation. With some glasses there is a good deal of "creep" after the initial instantaneous deflection. It is slowly recoverable, perhaps

TABLE 6. EFFECT OF SURFACE CONDITION

Glass rods 1/4 in. diam.
Load duration – 1 hr.

Treatment	*Breaking Stress, Psi*
Severely sandblasted	2,000
As received from factory	6,500
Acid etched and lacquered	250,000

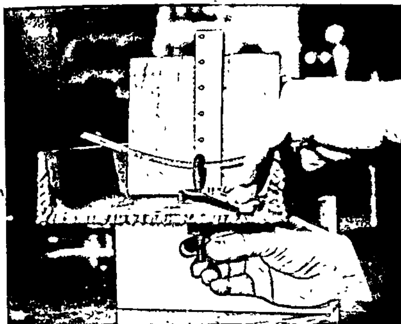


FIG. 30 Etched ½ in glass rod loaded to 125,000 psi by Phillips in 1936, as it looked in 1962 after 26 years under continuous load

completely, certainly largely, and so is not a permanent distortion But sometimes this recovery may last for a very long time.

Table 6 makes it obvious that deliberate mechanical damage to the surface can badly weaken the glass And, since etching in hydrofluoric acid actually removes some of the surface, it is equally obvious that surface must already have contained something which drastically weakened it Shand[15] points out that most flaws exist for a good reason For example, *it has* been known for many years that grinding cracks in plate glass and optical glass persist after polishing and can be revealed by a light hydrofluoric acid etch and suitable illumination. The originating flaws in ordinary plate glass can often be clearly seen under a magnification of 200×, and



FIG. 31. W. W. Shaver, under whose supervision the original work was done, has removed the load and the rod shows no observable permanent deformation.

Shand has measured depths from less than 0.0001 in. to as much as 0.005 in., a ratio of 1 to 50.

Statistical Theories

The flaw theory inevitably leads to a host of statistical theories of fracture. All are based on the reasoning that "a chain is only as strong as its weakest link," so they tend to fall into a pattern that has been summarized by Epstein.[36] All predict certain results which appear to be true under specific circumstances. We must not be surprised, for example, if we break a group of a hundred supposedly identical ¼ in. diam. commercial glass rods and get a spread between

maximum and minimum breaking stresses of 2.5 to 1. Statistically, this would be interpreted as meaning that very severe flaws were present in some rods, much less severe in others. These theories also suggest that strength should decrease as area under maximum stress increases, because, as area increases, so does the chance of finding a more damaging flaw. Thus, all other things being equal, glass tested in straight tension should be weaker than in cross-bending, and short glass rods, fibers, or laths should be stronger than long ones. Small-diameter fibers should be stronger than rods. There is a tremendous mass of evidence, accumulated over the past forty years, that these statistical predictions are correct when the glass exhibits strengths at the usual commercial level or strengths certainly not exceeding some such figure as 500,000 psi. In fact, a recent paper by Spriggs et al.[27] shows an area effect for other ceramic materials, including sapphire, MgO, and various grades of Al_2O_3 and SiC. The ratio in strength between bending and straight tension data is almost exactly 2 to 1, as it often is for glass.

However, when the glass being tested is very strong, the statistical theories fail in two ways. Although he did not report it until ten years later, Otto[3] had already found that, if drawn under the proper conditions, fibers could have equal or very nearly equal strength even when the ratio of diameters was as great as 2 to 1. Thomas,[5, 6] working with the same glass used by Otto, but with further refinements in technique, found no change whatsoever in strength for fiber diameters between 2 and 6×10^{-4} in. The prediction that strength should decrease as area increases is not fulfilled. Moreover, the scatter of results around the median value of 530,000 psi was only 1 per cent, whereas statistical theories predict that the dispersion must increase as the median failure strength increases. Mould,[8] reporting on work done somewhat earlier, broke fibers in cross-bending and also obtained moduli of rupture around 500,000 psi under loads

of several seconds' duration. For loads of 0.01-sec duration the strength increased to 610,000 psi. The scatter was approximately 2.5 per cent in both cases. Braithwaite and Sucov[12] worked with the same glass composition used by Otto and Thomas but with rather coarse fibers, 4×10^{-3} to 6×10^{-3} in. in diameter. The glass was melted by induction heating and the fibers were drawn and tested in a controlled, ultraclean atmosphere. The observed strengths were essentially constant at 405,000 psi ±2 per cent. Evidence of this type leads us to view the statistical theories with increasing caution. They are probably meaningful only where there are large numbers of flaws. This probably is not true for most fibers. Chechulin,[13] in fact, claims to show that the strength of small specimens must be virtually independent of size, and that, for fibers, this will be true for diameters below about 2×10^{-3} in. He claims that for thicker fibers the strength will depend not only on the degree of drawing but also on the diameter, to an extent that increases with the thickness of the fiber. There may be some degree of confirmation of these suggestions in a comparison of the results of Thomas on fine fibers and those of Braithwaite and Sucov on fibers ten times larger.

Standardization of Test Conditions

In an earlier paragraph, the expression "all other things being equal" was used. Unfortunately, with glass, other things are not always equal because the strength also depends on temperature, rate of stress application, atmosphere surrounding the specimen, and previous thermal history. Lack of appreciation and control of these variables has led to much discordant data in the literature. As Greene[14] says, "it is obvious that strength is not a well-behaved, reproducible property like density, thermal expansion, or viscosity." Preston[15] comments that some of the earlier approaches had "all the finesse of a bull charging a haystack."

Several things can be done to help standardize the test conditions. In actual service the glass will be subjected to a wide variety of man-made scuffs, scratches, and abrasions. Thus, a *standard* abrasion of controlled severity on the area of greatest stress in the test samples may give a surface more closely representative of conditions in actual use. At the same time, it will reduce the spread of test values. Of course, all hope of getting really high strengths is lost by this procedure.

A second method of controlling some of the variables depends on the fact that, in most tests, glasses exhibit static fatigue. In a very short time impact test of perhaps 10^{-4} sec, most glass articles are nearly twice as strong as in a test of several minutes' duration. Phillips[46] has shown that the same effect is exhibited by polycrystalline materials such as steatite, zircon, wollastonite, and three alumina compositions Furthermore, the effect is quantitatively the same, within ±9 per cent, as for a borosilicate glass On the other hand, if the glass is tested in very dry air, or in high vacuum, or for very short intervals (in each case after thorough outgassing), or at very low temperatures, such as that of liquid nitrogen, it shows no static fatigue at all Its strength is independent of rate of loading or time The actual numerical values of the strengths obtained in these several ways still depend upon the nature and extent of the surface damage, but all tend toward the *same* value and this is a maximum for that particular kind of glass and glass surface This, then, is a second way of standardizing the test conditions, and one that gives us a good deal of insight into some of the mechanisms involved in the fracture process

Static Fatigue

The most complete and systematic investigations of static fatigue are those of Preston,[47] Baker and Preston,[48] Mould and Southwick,[32] and Mould.[41, 49, 50] Mould studied strength as a function of five variables time, temperature, ambient

humidity, abrasion depth, and abrasion age. One of the most intriguing aspects of this work was the discovery of the "universal fatigue curve." The specimens to be discussed here were aged for 20 to 24 hours while immersed in distilled water and were broken at various time durations thereafter while still under water. The specimens were standard laboratory microscope slides which were abraded in six different ways and then broken in cross-bending so that all fractures originated in the abraded area. Twenty specimens were used in each test. Despite the variety of abrasive treatments, when reduced strengths (observed strength/liquid nitrogen strength) are plotted against reduced time (observed time/time at which strength has half its liquid nitrogen value), the remarkable result shown in Figure 32 develops, in which all the data are adequately represented by a single smooth curve.

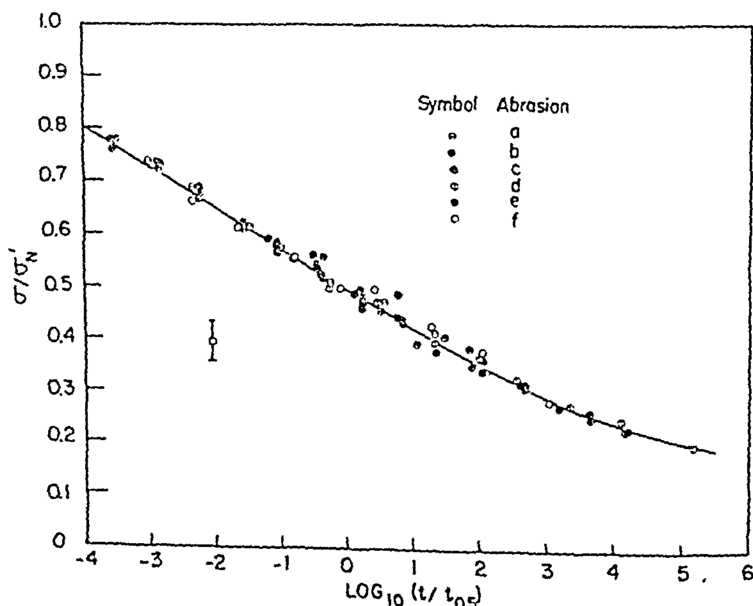


FIG. 32. Universal fatigue curve for glass abraded in various ways (after Mould). See text for details.

Mould has shown that the strength of *aged* abraded rods broken in liquid nitrogen may be as much as 60 per cent higher than that of freshly abraded rods. This effect was greatest when the rods were aged in liquid water or in humid atmospheres, but was entirely absent when they were stored in dry air or vacuum. Because of this, the curve at the long time side of Figure 32 should probably be split into a family of curves approaching various reduced strength values almost horizontally. At these durations, under certain conditions, perhaps largely controlled by temperature and glass composition, *the aging or healing of abrasions with time becomes* comparable to the static fatigue effect, and compensates for it. An "endurance limit" or limits will be observed so that there will be an applied stress below which breakage will not occur, regardless of the duration of the test. Mould agrees that the statement often found in the literature that the very long time strength is about one third of the very short time strength is approximately correct.

Mould believes that static fatigue results from changes, under stress, in the shape and/or depth of surface flaws which were developed by previous mechanical contact with the surface. *These changes* take place only when a surrounding medium can react with the newly formed surfaces, thus decreasing the surface energy. Water appears to be the most effective agent, and the rate of crack growth appears to be governed by the localized stress at the crack tip and the water or water vapor concentration there. Mould believes that the reaction is a simple hydrolysis involving only the water and the silica network. In contrast to ordinary chemical attack on glass surfaces, the presence of other ions in the glass or in the attacking medium seems to have little effect. It remains a moot question whether the effect is purely physical,[51] or purely chemical as in the stress corrosion theories of Charles[50] and of Charles and Hillig,[52] or perhaps a mixture of the two.

Four Ranges of Glass Strength

Some years ago Preston[45] suggested that there are really three ranges of "strength" that have been investigated in glass. Today we would have to say there are four subdivisions, with of course some overlapping.

(I) In the low strength range are most massive commercial articles: plate and window glass, containers, and the like. In almost every case, these will break at stresses ranging from 2,000 to 20,000 psi.

(II) In the second range, from perhaps 30,000 to 100,000 psi, are commercial articles which have been strengthened by placing the surfaces in compression. This may be done by physical means including chill tempering, or by chemical means (chemical tempering) including high and low temperature ion exchange and surface or volume crystallization.

By controlled heating and cooling, using air under pressure, we can make glass articles two to four times stronger than they would normally be. All exposed surfaces are in permanent compression with the compensating tensile stresses buried inside where they can do no harm. Since glass always breaks in tension and from the surface, these surface compressive stresses must first be overcome, and the glass is thus effectively strengthened. This kind of glass is called "chill tempered" or "toughened." You see it, but probably do not recognize it, in the side and rear windows of your automobiles. Chill tempering has some rather severe limitations in respect to the size and shape of the glass that can be treated. Corning Glass Works has recently developed other treatments called "chemical tempering," which place the surfaces in even higher compression. In this way strengths up to 100,000 psi are obtained. Even when abraded, some of the chemically tempered glasses can have a modulus of rupture above 50,000 psi. When compared on an equal weight basis, these glasses are about equivalent in strength to 410 stainless steel.

(III) In a third category, ranging roughly from 100,000 to 500,000 psi, are some of the fiber glasses, other glasses which have received very special care in manufacture, and still others which have been acid etched. Thomas (as reported by Jellyman[19] and Holloway,[53] found that freshly drawn rods could support stresses of over 500,000 psi in the regions which were free from visible defects Proctor,[54] working with rods which were not touched in the stressed areas between removal of the molten glass from the furnace and the final fracture, obtained an average unetched strength of 152,000 psi with a maximum of 387,000 psi When etched, the strengths sometimes exceeded 500,000 psi, but there was considerable scatter in the results.

The scatter just mentioned seems to be characteristic of most of the earlier results with acid-etched glass. Etch pits and deposit marks are often observed, but their removal does not seem to improve the strength significantly. Recently, however, Brearley et al,[55] using commercial rods heated and drawn down to diameters of 0 012 to 0 036 in, have reported some remarkably *consistent* results with etched glasses of three compositions After two hours of etching in dilute hydrofluoric acid, a soda-lime glass reached a plateau of 150,000 psi, a borosilicate 380,000 psi, and a lead glass 300,000 psi Carefully hand-drawn but *not* etched soda-lime and borosilicate glass samples had average strengths 20 to 25 per cent higher than those listed above, but the scatter was much greater. The same two glasses had still higher strengths when *etched* and then broken under liquid nitrogen. Strangely enough, the borosilicate now became the stronger of the two, rising to an average of 710,000 psi, while the soda-lime rose only to 625,000 psi It is obvious that there is much still to be learned in this area, and it is equally obvious that we have already transgressed into strength subdivision (IV)

Before leaving this section, the work of Symmers et al[56] should be mentioned They worked with 4-mm soda-lime

rods heated and necked down to 0.020 in., and broken in straight tension. They "standardized" these rods by giving them a 30-sec dip in 30 per cent hydrofluoric acid to clean the surfaces. They then had an average strength of 77,000 psi ±7 per cent. When etched, they did not reach some of the very high strengths found by other workers, but they could get to 250,000 psi. What is especially significant is that this strength could be reduced 25 per cent simply by rubbing the rods with a coarse cloth. There was about the same reduction when one rod was drawn across another with a "sawing action." But there was a 70 per cent loss in strength when only 2.5 grams of ground and sieved glass was dropped 1/4 in. onto the etched surface! The rods were then only very slightly stronger than the "standardized" glass.

(IV) In this category is a variety of observations ranging from roughly 500,000 psi to well over 2,000,000 psi. The remarkably consistent results of Thomas with fine fibers have already been mentioned. Hillig[57] used larger rods, about 1 mm in diameter, drawn down in an oxyhydrogen flame from large commercial stock. No special storage was employed except to avoid mechanical contact. Samples several months old gave the same results as newly prepared rods. The whole bending apparatus was immersed in liquid nitrogen. Breaking stresses ranged all the way from 100,000 psi to 1,970,000 psi for a sample 0.022 in. in diameter. The distribution was bimodal, with nine samples breaking at about 750,000 psi and ten others at 1,150,000 psi. Still lower temperatures (liquid helium and liquid hydrogen) had no effect on the strengths observed.

Morley et al.[58] worked with much smaller fused silica fibers, about 0.001 in. in diameter, prepared and tested under well-defined conditions. They also found a wide range of strengths which decreased in this order: liquid nitrogen immersion; vacuum at room temperature; air at room temperature; and air at higher temperatures. There was evidence

of static failure even in these small fibers. In liquid nitrogen, the strengths ranged from 1,700,000 to 2,300,000 psi. There was no evidence of a bimodal distribution. They concluded that their samples still contained flaws but did not speculate further.

Provance[59] and others have reported fiber strengths ranging from 600,000 to over 1,000,000 psi, when quickly broken in air. These results are for commercial diameters of 0.0002 to 0.0005 in.

Finally, there are the very high strengths reported for "whiskers." Probably because of their very good surfaces, no further preparation is needed to make them very strong. Brenner[60] reports a value of 2,200,000 psi for Al_2O_3 and Ryshkewitch[61] gives 2,800,000 psi for BeO.

Current Status of the Microcrack

The impact of the Griffith crack theory was so great that it grew and prospered over the years despite the fact that no one had ever really seen such microcracks. Many methods, including optical and electron microscopy, were employed without success. Man-made scratches of course are visible, but even for glass of quite modest strength the surface is substantially featureless.

Andrade and Tsien[62] reported that hard borosilicate glass, exposed to sodium vapor at 350°C or above, showed a network of fine cracks. The paper was widely discussed, but it seemed that the treatment was so severe and the temperatures so high that the interpretation of these cracks as corresponding to already existing cracks was quite questionable.

After a fitful slumber of twenty years, four laboratories in rapid succession reported additional work with sodium vapor: Gordon et al.,[63] Argon,[64] Nakayama,[65] and Ernsberger.[66] The overall conclusion appeared to be that the cracks in the Andrade-Tsien pattern that are artifacts may indeed originate

at true microcracks. Ernsberger suggested that, to be more certain of this, a method was needed which would avoid high temperatures and which would put only the surface in a two-dimensional state of tension. Fortunately, such a method exists. It is called ion exchange and has been mentioned previously in connection with chemical tempering, although compression rather than tension was desired there.

Ion Exchange

Soda-lime glass can act as a cation exchanger. Most, if not all, of the sodium ions are exchangeable, on a one-for-one basis, with other monovalent ions. The exchange obeys Fick's law, and the diffusion constant is determined by temperature and by the size of the larger of the exchanging ions. If the soda-lime glass is immersed in a molten bath containing lithium and potassium nitrates, the lithium ions will begin to replace the sodium. The potassium salt is nearly inert, in this case, so far as ion exchange is concerned, but has the highly desirable effect of lowering the melting point of the mixture. As the sodium ion is replaced by the smaller lithium, the oxide network tries to shrink into a smaller space. Since the exchange is limited to a very thin surface zone (perhaps 10^{-4} in. deep), the shrinkage can be only perpendicular to the surface, the massive glass underneath preventing shrinkage in the other two directions. The thin surface layer is then under hydrostatic tension in two directions and, if flaws are present, fracture will occur when the tension becomes great enough.

One of the advantages of the lithium salts over the ammonium bisulfate used by Acloque et al.[67] is that a molten droplet will not spread on glass. In his truly elegant experiments, Ernsberger[66] took advantage of this property to avoid the flaws known to be associated with all cut edges of glass. He was thus able to show that the number of crack

origins revealed by ion exchange does *not* increase indefinitely with time of treatment. In fact there are areas in acid-etched, fire-polished, or freshly cleaved surfaces that are entirely free of crack origins no matter how prolonged is the treatment. Another useful feature of this technique is that, when time and temperature are suitably chosen, the crack growth does not occur during the ion exchange itself, nor while the glass is cooling, but only after the treated surface is dipped into water.

Further refinements of this process,[illegible] by a photoresist technique, permitted ion exchange in areas small enough that an accurate crack count could be made. In glazing-quality plate glass there appeared to be, on the average, 35,000 cracks per square inch which presumably are fissures produced in grinding and which have not been eliminated by the polishing operation. In mirror-quality plate glass, the number was reduced to 4,500 per square inch. In window glass, the flaws were much fewer in number and apparently are the result of mechanical damage in making and handling. For both plate and window glass, etching in dilute hydrofluoric acid sharply reduces the number of microcracks. Few survive the removal of 5μ, but etch marks remain to mark their original positions.

The conclusion that mechanical damage leads to microcracks seems certain. This need not be the kind of damage visible to the naked eye or even to the electron microscope. In one experiment, Ernsberger drew a flexible glass fiber lightly across a clean, dry, hydrofluoric acid-etched surface. No damage was visible optically, but ion exchange revealed an extensive, complicated, herringbone pattern of microcracks. In another experiment, 320-mesh abrasive particles were rolled between two sheets of hydrofluoric acid-etched glass. The ion-exchange pattern is shown in Figure 33. The work of Symmers, Ward, and Sugarman should be recalled in the context of these results.

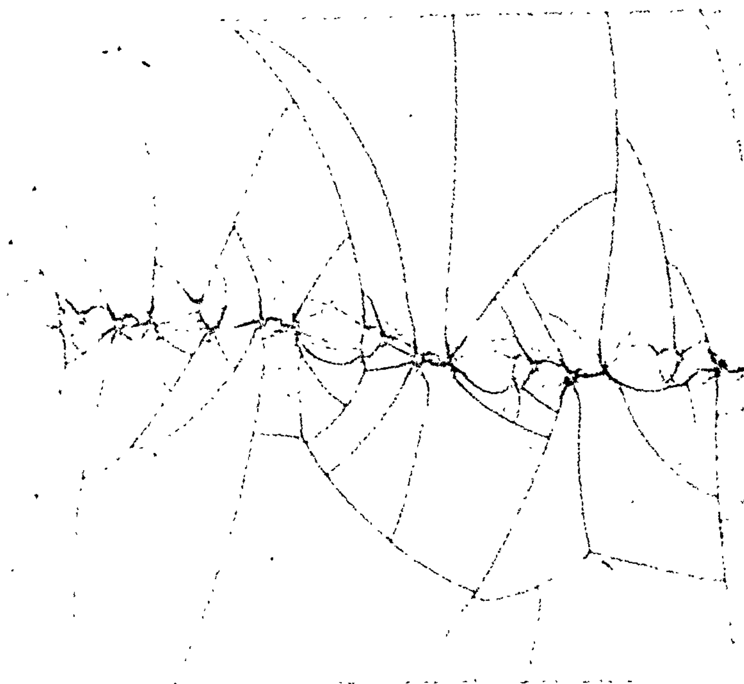


FIG. 33. Ion-exchange crack patterns reveal surface damage produced by abrasive particles rolled across initially flawless surface (115×) (after Ernsberger).

Ernsberger also believes that devitrification nuclei can develop at temperatures far below those where devitrification would seem to be possible, and that these nuclei can be the source of surface cracks. Figure 34 shows the result when commercial sheet glass was heated for 2 hr at 650°C, ion exchanged, and then etched lightly. He also shows that an inverse relationship exists: microcracks can nucleate devitrification.

Pulsed Stress Experiments

The microcrack population on glass surfaces is normally invisible because the crack surfaces are in optical contact.

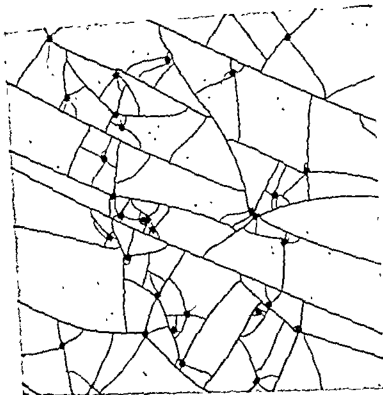


FIG. 31 Ion-exchange pattern showing strength-impairing effect of surface devitrification (105×) (after Ernsberger)

When the one worst flaw propagates rapidly the sample breaks, but none of the other cracks becomes visible, because the microscopic failure has relieved the stress on them If the stress were of sufficiently short duration one might expect to limit the growth of the most serious microcrack and make it visible without destroying the specimen If a very short time pulse could be made to travel through the sample, perhaps many microcracks might be made visible J E Field[69] has done this A drop of water, traveling at supersonic speed, was made to strike a glass plate etched with HF up to a boundary line The impact was made to occur at the boundary It was found that visible cracks occurred only in the unetched portion of the field

Two radically different types of experiments (ion exchange and pulsed stress) thus demonstrate the existence of invisible microcracks on the surface of glass. But are *all* pre-existing flaws activated to grow and become visible under these test conditions? We do not know.

Preliminary Conclusions and New Problems

Optically undetectable cracks can now be located, counted, and associated with several types of mechanical damage. The initiating flaw, after fracture, can also be measured optically within strength ranges I and II. Perhaps no other types of flaw-making mechanisms are necessary to account for strengths below approximately 500,000 psi. The new frontier has therefore shifted to apparently undamaged surfaces such as on the fire-polished fused silica rods which Hillig found to be flaw-free under the electron microscope. These still display strengths two to five times lower than theory predicts; with one or two exceptions, strengths show a wide scatter; and static fatigue still exists.

Are there other, "intrinsic" flaws? Do tiny, invisible bubbles act as stress concentrators? Is there glass-in-glass phase separation on a submicroscopic scale, as Watanabe et al.[70] and many others have suggested? Are the phase boundaries in such a system planes of weakness? Is this perhaps why fused silica fibers, presumably free of phase boundaries, seem consistently stronger than fibers from other glasses? And what about thermal history—is it or is it not important on this high-strength plateau? It is obvious that there are still as many questions as there are answers. Let us therefore turn to some new data and to some new interpretations of old data.

Calculation of Young's Modulus from Composition

With one exception, all of the force–separation assumptions, such as the simple one that resulted in equation (4),

show that the ultimate strength σ_u should be directly proportional to Young's modulus E. It should thus be of paramount interest to know how E varies with the chemical composition of different glasses. Phillips[71] describes a method of doing this for both simple and complex silicate glasses. The content of each of the oxides is expressed in mole per cent, such as p_1 for SiO_2, p_2 for Al_2O_3, etc. Each oxide is also assigned a numerical coefficient such as C_1 for SiO_2, C_2 for Al_2O_3, etc. The modulus of elasticity is the sum of the terms $C_1p_1 + C_2p_2 + \ldots C_np_n$. This is not quite the simple linear relationship it appears to be. Some of the coefficients, such as those for the alkalis, depend on and vary with the total content of SiO_2 plus Al_2O_3 plus a fraction of the B_2O_3. Others vary with the alkali content. One—the coefficient for B_2O_3—depends on the relative amounts of Al_2O_3 and alkali with which it is associated. Data from thirty-six glasses reported in the literature are used to develop some very simple and systematic computational methods. When these are applied to thirty-five *different* glasses, some with as many as eight constituents, the average difference between calculated and observed values of E is less than ± 0.3 per cent, with the extreme differences being -0.5 and $+1.0$ per cent. These deviations are an order of magnitude less than those observed by earlier workers.

In this method of computing E, the SiO_2 coefficient is 7.3 kilobars per mole per cent for all glasses, including fused silica for which Spinner[72] found E to be 730 kilobars. This use of 7.3 is perhaps tantamount to saying that, regardless of whatever other constituents are present, the SiO_2 retains its identity in the structure, at least so far as elasticity is concerned. It does *not* mean, as Charles[8] has suggested by rather indirect evidence, that E depends almost entirely on the Si–O–Si linkages alone. On the contrary, the effects of the Si–O–Na bonds are shown to be far from negligible.

In the Phillips procedure, the coefficient for CaO is high

and constant at 12.6 kilobars per mole per cent. Al_2O_3 also has a coefficient with a high value, and direct observation shows us that some of the highest E values observed are in calcium aluminosilicates. Obviously much remains to be learned in this area.

Dependence of Strength on Young's Modulus

Despite the inadequacies of equations such as (4), a number of people have commented, usually rather casually, that *observed* (not ultimate) strengths do often seem to be proportional to Young's modulus. Few seem to realize that, at commercial strength levels, this is more often the rule than the exception.

Data for glasses. Schwalbe et al.[73] broke coarse fibers, 0.024 to 0.032 in. in diameter, in straight tension, using five different glass compositions. The work was done one hour after annealing was completed. The compositions fall within the area where the Phillips method can be used to compute E with considerable accuracy. When the breaking strengths σ_a are plotted against E, the straight line marked S, B, S in Figure 35 is the result. Working with 1/4 in. diam. rods "as received" from the factory, Phillips, measured the transverse strengths shown on the line marked P in Figure 35.[74] Watanabe, Caporali and Mould[75] worked with twenty one glass compositions. All were made into rods 0.050 to 0.085 in. in diameter, were given a standard abrasion, and were then quickly broken in four-point bending in liquid nitrogen *without* any aging. They themselves reject glass 21 because of its anomalous behavior. Glasses 18 and 19 are outside the limits of the Phillips method of calculating E. The others can easily be calculated and range between the limits of 9,400,000 and 12,600,000 psi. The average deviation in the strength values of the eighteen remaining glasses was ±9.0 per cent. When plotted against E, only two of the observed strengths

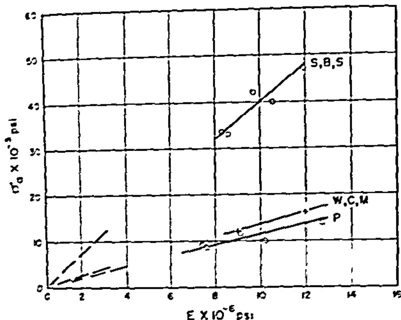


FIG. 35. Linear relationship between strengths and Young's moduli of glasses

are more than 9 per cent away from the straight line marked W, C, M on Figure 35, and these two are very close. On the other hand, when the glasses are aged over water, or actually broken in water, there appears to be absolutely no correlation with *E*. Now, these are all fairly low-level strengths, but there is evidence that the same correlation between σ_a and *E* exists for fibers with strengths ranging from 500,000 to 800,000 psi and with moduli between 10,700,000 and 19,000,000 psi. Figure 35 shows that doubling *E* also doubles σ_a. The same seems to be true for fibers, although the magnitude of the change is much greater.

At this point a simple modification of the Griffith equation (5) might be interesting. If we suppose that, in the handling all glass receives in manufacture (or under controlled abrasion), the actual damage as reflected in the crack depth *a* is inversely proportional to *E*, a not ...

can write $a = d/E$, where d is simply a proportionality constant and equation (9) then becomes

$$\sigma_u = E\left(\frac{2S}{\pi d}\right)^{1/2} \tag{11}$$

This reflects dependence of σ_u on E, and can also account for different magnitudes of σ_u through changing values of d and/or S.

Data for Polycrystalline Ceramics

Smoke[76] suggested in a qualitative way that those polycrystalline ceramics with high E values would also have high tensile and transverse strengths and vice versa. Davis and Clements in Walton[77] showed that σ_u/E was quite constant for certain types of refractories and aluminosilicate bricks. Fenstermacher and Hummel[78] measured E and transverse strength on thirteen specimens of $2Al_2O_3{\cdot}SiO_2$ mullite bodies. Although E was not a constant at room temperature, neither was σ_u, and a plot of σ_u versus E was a straight line with $\sigma_u/E = 10^{-3}$ on the average. Phillips[46] broke ½ in. diam. commercial ceramic rods (all supposed to be fully vitrified and nonporous) in three-point bending with a 4 in. span and a constant loading rate of 10,000 psi/min. His results are shown on Figure 36. There is the usual scatter, but again $\sigma_u/E = 10^{-3}$. Only the extreme values of the Fenstermacher and Hummel data are shown, but they fall only slightly above and slightly below the Phillips line. The point at $E = 10^6$ psi and $\sigma_u = 11.7 \times 10^3$ psi is for a borosilicate glass and is quite close to the line representing the glasses tested many years before.[71] Finally, five points representing catalog data are shown. These are flexural strengths with the tests performed in accordance with A.S.T.M. standard D116-61T. The loading rate in that test is considerably more rapid than that used by Phillips. To take this into account the catalog strengths

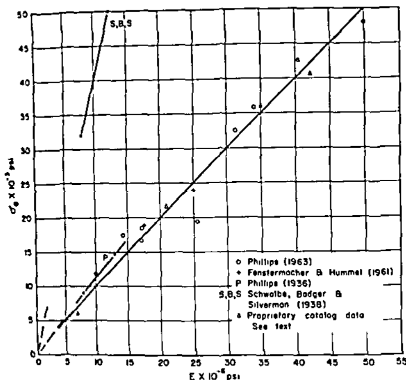


FIG. 36 Linear relationship between strengths and Young's moduli of polycrystalline ceramics

have been multiplied by 0.9. They then fall very close to the major straight line. Omitting the higher strengths represented by the Schwalbe, Badger, and Silverman line where $\sigma_o/E \sim 4 \times 10^{-3}$, and the older Phillips data, this major line shows the data for six alumina ceramics (with 85 to 99 per cent Al_2O_3 content), two grades of steatite, two grades of zircon, and one each of cordierite, mullite, wollastonite, and borosilicate glass.

For polycrystalline ceramics, Hasselman[79] has already suggested that the "determination of Young's modulus may constitute an indirect non-destruction test for the effect of porosity on strength." The present data go farther and sug-

gest this is also true for essentially nonporous bodies and that the same or very similar strength-impairing mechanisms must exist for all brittle materials, whether amorphous or crystalline.

Re-examination of Inglis and Griffith Criteria

A final re-examination of the Inglis and Griffith criteria may now be in order. In equation (8) we still do not know what a realistic minimum value for the tip radius should be. Although the existence of microcracks now seems firmly established in many cases, we do not know their actual length, width, or depth. It is hoped that such data may become available in the near future, but at the moment we do not know what either a or ρ should be in the Inglis equation. Furthermore, if the cracks are numerous, as they often seem to be, we would expect some interaction between them, whereas the Inglis relationship, and all statistical theories, assume that each flaw acts independently.

Examination of the Griffith criterion, equation (9), leads to additional questions. E appears, but what value of E should be used? Hillig[57] has shown quite clearly that for fused silica the numerical value of E increases with stress and is 60 per cent greater at a stress of 2,000,000 psi than at zero stress. The relationship between σ_a and E is nearly linear, but the departure from linearity is in the direction *opposite* to that predicted by Figure 27. If all glasses and, in fact, all brittle materials behave in this way, then E will have to be made stress-dependent.

Another question-mark is the surface energy S. Data on surface energies of solids are meager. Only recently has reliable information been obtained for copper. With nonmetallic solids, and particularly those with mixed ionic-covalent bonding, as in glasses, the problem becomes extremely complex. Yet Griffith extrapolated from 2000°F to

room temperature! Bikerman[40] offers an even more serious objection. He does not believe that the surface energy S is involved at all in fracture phenomena. He argues that work is done, not on rupture as such, but on the deformations leading to rupture. Whereas the Griffith theory considered a competition between strain energy and surface energy, the Bikerman concept requires only a redistribution of strain energy. Whereas the old theory assumes that none of the work of fracture becomes heat, the new theory assumes that most of it will be converted to heat. The experimental results (Kuznetsov,[41] Schellinger[42]) appear to be in better agreement with the new than with the old hypothesis. The specific fracture work F, to produce one cm² of new crack surface, then becomes several times greater than the S value of Griffith. Shand[43] has also discussed how greatly S itself can vary. Bikerman finally arrives at an equation

$$\sigma_a = \left(\frac{EF}{2a}\right)^{1/2} \qquad (12)$$

which is similar in most respects to the Griffith relationship but in which F is much larger than Griffith's S.

If we modify equation (12) by again making $a = d/E$, we can write

$$\sigma_a = E\left(\frac{F}{2d}\right)^{1/2} \qquad (13)$$

Some method of further investigating equations (11), (12), and (13) would certainly appear to be a most worthwhile enterprise.

As an alternative to the Griffith theory, Poncelet[44] has proposed that strength-controlling flaws are actually generated by the application of stress, through natural thermal fluctuations biased by the stress field. In mathematical form the resulting equations for size effect, crack propagation rate, strength variability, and static fatigue all seem to be reasonable. However, they are so similar to the relationships derived

from the Griffith theory that, as yet, there is no clear-cut way to choose between the two alternatives. A radically different approach is that of Marsh,[26, 27] who feels that the true brittle fracture of glass is limited to perhaps 500,000 psi, and that higher breaking stresses result from plastic (flow with a large increase of specific fracture energy) around the crack tip.

Summary

In the fifty years which have passed since the pioneering work of Inglis and Griffith, much progress has been made in understanding the important and tantalizing problems of the strength and weakness of brittle materials. However, much also remains to be learned. Perhaps some of the new concepts outlined here will eventually allow us to correlate chemical composition, structure, elasticity, and strength. There will no doubt still be the surprises, contradictions, and frustrations already outlined. But hope will also remain at the bottom of Pandora's magical box.

The author wishes to express his sincere thanks to many individuals and companies for their assistance: to J. H. Munier, C. J. Parker, and W. W. Shaver, all of Corning Glass Works, and to that company itself for permission to publish certain old but useful data; to F. W. Ernsberger and Pittsburgh Plate Glass Co.; to J. J. Bikerman and Horizons Incorporated; to R. E. Mould and American Glass Research, Inc.; to F. W. Preston and E. B. Shand, consultants; and to H. T. Smyth, Rutgers, The State University.

Appreciation should also be extended to the Selenium-Tellurium Development Association, Inc. and to the International Lead Zinc Research Organization for financial support of parts of this work.

REFERENCES

1. Dora and Erwin Panofsky, *Pandora's Box*, Pantheon Books, New York (1956), p. xiii
2. G. O. Jones, *J. Soc. Glass Technol.*, *33*, 120 (1949).
3. E. Orowan, *Rep. Progr. Phys.*, *12*, 185 (1949)
4. E. B. Shand, *J. Am. Ceram. Soc.*, *37*, 52 (1954).
5. W. H. Walton, *Mechanical Properties of Non-Metallic Brittle Materials*, Interscience, New York (1958)
6. B. L. Averbach et al., *Fracture*, Technology Press, Massachusetts Institute of Technology and Wiley, New York (1959)
7. R. J. Charles and J. C. Fisher, in *Non-Crystalline Solids*, V. D. Frechette, Ed., Wiley, New York (1960) p. 191
8. R. J. Charles, *Progress in Ceramic Science*, Pergamon Press, New York *1*, 1 (1961)
9. W. B. Hillig, in *Modern Aspects of the Vitreous State*, J. D. Mackenzie, ed., Butterworths, Washington (1962), p. 152
10. F. M. Ernsberger, in *Progress in Ceramic Science*, J. E. Burke, Ed., Macmillan, New York, *3*, 57 (1963)
11. S. Schurkow, *Phys. Z. Sowj.*, *1*, 123 (1932)
12. S. Schurkow, *Tech. Phys. USSR*, *1*, 386 (1935)
13. L. C. Lyle, in *Mechanical Properties of Non-Metallic Brittle Materials*, W. H. Walton Ed., Interscience, New York (1958) p. 116
14. S. W. Bradstreet, *Bull. Am. Ceram. Soc.*, *37*, 510 (1958)
15. A. A. Griffith, *Phil. Trans. Roy. Soc.*, *221*, 163 (1920), also *International Congress for Applied Mechanics* (1924) p. 61
16. W. H. Zachariasen, *J. Am. Chem. Soc.*, *54*, 3841 (1932)
17. B. E. Warren, *Z. Kristallogr.*, *86*, 349 (1933)
18. I. Prodhomme, *Glass Ind.*, *39*, 587 (1958)
19. P. E. Jellyman, in *Fibre Structure*, Textile Institute and Butterworths, London (1963), p. 567
20. J. J. Gilman, in *Mechanical Behavior of Crystalline Solids*, NBS Monograph, Washington, *59*, 79 (1963)
21. I. Naray-Szabo and J. Ladik, *Nature*, *188*, 226 (1960)
22. C. E. Inglis, *Trans. Inst. Nav. Archit.*, London, *55*, 219 (1913)
23. H. A. Elliott, *Proc. Phys. Soc.*, *59*, 208 (1947)
24. C. Gurney, *Phil. Mag.*, *39*, 71 (1948)
25. E. F. Poncelet, *Verres et Réfr.*, *5*, 69, in four parts (1951)
26. D. W. Marsh, *Fracture of Solids*, Interscience and Wiley, New York (1963)

27. D. W. Marsh, *Proc. Roy. Soc.*, *279*, 420 (1964).
28. E. Orowan, *Weld. J.*, London, *34*, 1575 (1955).
29. J. P. A. Tillet, *Phys. Soc. Lond. Proc.*, B., *69B*, 47 (1956).
30. R. J. Charles, *J. Appl. Phys.*, *29*, 1554 (1958).
31. W. C. Levengood, *J. Appl. Phys.*, *29*, 820 (1958).
32. R. E. Mould and R. D. Southwick, *J. Am. Ceram. Soc.*, *42*, 542, 582 (1959).
33. E. B. Shand, *J. Am. Ceram. Soc.*, *44*, 21 (1961).
34. J. P. Berry, *J. Polymer Sci.*, *50*, 107, 313 (1961).
35. E. B. Shand, *J. Am. Ceram. Soc.*, *47*, 102 (1964).
36. B. Epstein, *J. Appl. Phys.*, *19*, 140 (1948).
37. R. M. Spriggs, L. A. Brissette, and T. Vasilos, *Mater. Res. Std.* (May 1964), p. 218.
38. W. H. Otto, *J. Am. Ceram. Soc.*, *38*, 122 (1955).
39. W. F. Thomas, *Nature*, *181*, 1006 (1958).
40. W. F. Thomas, *Phys. Chem. Glasses*, *1*, 4 (1960).
41. R. E. Mould, *J. Appl. Phys.*, *29*, 1263 (1958).
42. D. E. Braithwaite and E. W. Sucov, paper 2-G-63, 65th Annual Meeting, American Ceramic Society (1963).
43. B. B. Chechulin, *Zhur. Tekh. Fiz.*, *24*, 292 (1954).
44. C. H. Greene, *J. Am. Ceram. Soc.*, *39*, 66 (1956).
45. F. W. Preston, *Bull. Am. Ceram. Soc.*, *33*, 355 (1954).
46. C. J. Phillips and S. Di Vita, *Bull. Am. Ceram. Soc.*, *42*, 685 (1963).
47. F. W. Preston, *J. Appl. Phys.*, *13*, 623 (1942).
48. T. C. Baker and F. W. Preston, *J. Appl. Phys.*, *17*, 162 (1946).
49. R. E. Mould, *J. Am. Ceram. Soc.*, *43*, 160 (1960).
50. R. E. Mould, *J. Am. Ceram. Soc.*, *44*, 481 (1961).
51. E. Orowan, *Nature*, *153*, 341 (1944).
52. R. J. Charles and W. B. Hillig, Symposium sur la resistance mecanique du verre et les moyens de l'ameliorer, 1962; Union Scientifique Continentale du Verre, Charleroi, Belgium.
53. D. G. Holloway, *Phil. Mag.*, *4*, 1101 (1959).
54. B. Proctor, *Phys. Chem. Glasses*, *3*, 7 (1962).
55. W. Brearley, P. A. P. Hastilow, and D. G. Holloway, *Phys. Chem. Glasses*, *3*, 181 (1962).
56. C. Symmers, J. B. Ward, and B. Sugarman, *Phys. Chem. Glasses*, *3*, 76 (1962).

57. W. B. Hillig, *J. Appl. Phys.*, *32*, 741 (1961).
58. J. G. Morley, P. A. Andrews, and I. Whitney, *Phys. Chem. Glasses*, *5*, 1 (1964).
59. J. D. Provance, Paper presented at April 1964 annual meeting of the American Ceramic Society; see also *Chem. Week* (April 11, 1964), p. 59.
60. S. S. Brenner, *Growth and Perfection of Crystals*, Wiley, New York (1958), p. 157.
61. E. Rashkewitch, *Sci. Technol.* (Feb. 1962).
62. E. N. daC. Andrade and L. C. Tsien, *Proc. Roy. Soc.*, *A*, *159*, 346 (1937).
63. J. E. Gordon, D. M. Marsh, and Margaret E. M. L. Parratt, *Proc. Roy. Soc.*, *A*, *249*, 65 (1959).
64. A. S. Argon, *Proc. Roy. Soc.*, *A*, *250*, 472 (1959).
65. J. Nakayama, *J. Phys. Soc.*, *Japan*, *14*, 1107 (1959).
66. F. M. Ernsberger, *Phys. Chem. Glasses*, *1*, 37 (1960).
67. P. Acloque, P. LeClerc, and P. Ehrmann, "Compte rendu du Colloque sur la Nature des Surfaces vitreuses polies," Paris, 23-24 Nov. 1959, and Round Table on "Polissage du Verre," Brussels, 13-14 June 1960, Union Scientifique Continentale du Verre, Charleroi, Belgium.
68. (*a*) F. M. Ernsberger, *Proc. Roy. Soc.*, *A*, *257*, 213 (1962).
    (*b*) F. M. Ernsberger, *Advances in Glass Technology*, Plenum Press, New York (1962), p. 511.
69. J. E. Field, Symposium sur la resistance mecanique du verre et les moyens de l'ameliorer, 1962, Union Scientifique Continentale du Verre, Charleroi, Belgium.
70. N. Watanabe, H. Noake, and T. Aiba, *J. Am. Ceram. Soc.*, *42*, 593 (1959).
71. C. J. Phillips, *Glass Technology*, *5*, 216 (1964).
72. S. Spinner, *J. Am. Ceram. Soc.*, *37*, 229 (1954).
73. W. L. Schwalbe, A. E. Badger, and W. B. Silverman, *J. Am. Ceram. Soc.*, *21*, 333 (1938).
74. C. J. Phillips, unpublished work done at Corning Glass Works (1936).
75. M. Watanabe, R. V. Caporali, and R. E. Mould, *Phys. Chem. Glasses*, *2*, 12 (1961).
76. F. J. Smoke, *Ceramic Age*, reprint (1954).
77. W. R. Davis and J. F. Clements, in *Mechanical Properties of Non-Metallic Brittle Materials*, W. H. Walton, Ed., Interscience, New York (1958), p. 203.

78. J. E. Fenstermacher and F. A. Hummel, *J. Am. Ceram. Soc.*, *44*, 297 (1961).
79. D. P. H. Hasselman, *J. Am. Ceram. Soc.*, *46*, 564 (1963).
80. J. J. Bikerman, *Glass Ind.* (1963), p. 382.
81. V. D. Kuznetsov, *Surface Energy of Solids*, H. M. Stationery Office, London (1957), p. 160.
82. A. K. Schellinger, *Science*, *111*, 693 (1950).

8. THE LOGIC OF THE MIND

By J. Bronowski
Salk Institute for Biological Studies

I am honored to contribute to this series, under the auspices of two famous societies, and I am also and rather particularly grateful. I am grateful, of course, that you have asked me, a newcomer to the American scene, who is still at the delightful stage of being bewildered by the unexpected range and the energy of your scientific interests

Yet when I said that I am particularly grateful, I had a second reason in mind It is this A man of my preoccupations, that is, a practicing scientist with a passion for literature, is often asked to discuss the relation of science to literature. But usually he is constrained, by the broad nature of his subject and of his audience, to do so in very general terms. I feel no such constraint here This gives me the chance to write more searchingly and, as it were, more professionally about my subject than usual; and I take that chance gratefully

I gave a series of lectures recently at the American Museum of Natural History in New York on science and literature as modes of knowledge, which have been published as a book under the title *The Identity of Man* There are a few places here and there in the book (I think I count four) where I should have liked to speak more fully and more circumstantially, had I been speaking to a professional audience I shall single out one of these places, which sketches what may be called the machinery of the mind, and make it the occasion for a larger analysis here I hope to develop the others on other occasions.

The mind is an elusive entity, whose workings are not wholly confined within the brain. But because I am looking at the logical processes of the mind, it is fair that I concentrate in the first place on the brain as the organ in which these processes must be mechanized. The subject of how the mind works as a mechanism—what machinery we can imagine to operate within the brain—has its intrinsic interest in any case. We know that the brain is made of the same stuff as the rest of nature, and its atoms must therefore obey the same natural laws that other atoms do. In that sense, then, it is tempting and even reasonable to say that the brain must be some kind of a machine. But alas, to use the word *machine* in this catch-all sense misses the crux of the question. The real question about the human mind lies deeper. It asks: Is the brain a machine with a formal procedure of any kind that we can now conceive? Let me quote a pertinent passage from *The Identity of Man.*

> A machine is not merely a whirring train of gears or a humming set of electric circuits. These happy, busy strings of hardware are only the middle step, the visible link, in a procedure which has three steps, and to which the other two are as integral as this is. The machine is the procedure, and the whole procedure, all three steps of it. The first step is the instruction or input, which is the modern form of the button that starts the machine: and which must itself be precise and mechanical, an unequivocal set of holes or marks on a tape that directs the machine into one branch of its network of possible paths. Then comes the physical machinery which obediently carries out the instructions and turns them into actions. And the third step is the result or output, which is equally decisive and definite: in a computer, it is another set of holes or marks on a tape.
>
> It is of cardinal importance here, and essential to my

description, that the output from a machine must be exact and unambiguous as the input is. For a modern machine, like a man, is asked in part to regulate itself, and for this purpose it must be able to feed its output back into itself as a new instruction. Its output must therefore be as sharp, within the tolerance of the machine, as capable of symbolic expression, as well defined and as single-minded as its input.

Language of the Brain

Our field of inquiry is the gray region between the input to the brain and the output from it; that is, between the information that the senses send to the brain and the instructions or other decisions that issue from it. In this gray region the brain manipulates the input and draws conclusions from it. During this process, the brain presumably uses some symbolism which translates and codifies its conceptions of the outside world. We do not know what this symbolic language is, but if it is indeed to be mechanical (on any system that we understand), its units must consist of configurations of atoms and of changes in these configurations which are displayed as electrical signals. If then the brain reasons like a logical machine, these signs or units that it employs in its reasoning must constitute a formal language or series of languages which follow precise rules, just like the language of symbols in which we write out logical and mathematical arguments. The brain cannot be a machine in any sense that we understand unless its language is as strict and as artificial (in the logical sense) as any of our own marks on a magnetic tape.

The symbols with which the brain works—its language (or successive languages)—are physical, chemical and electrical. But this makes them no different from marks on a paper or on a tape. Provided that they are exact, and are always translated exactly in the same way, they constitute a formal

logical language. What is to be said about them, then, comes not from physics and chemistry and biology but from symbolic logic. This is why I, a mathematician, presume to talk about it to physicists and chemists and specialists in biology.

Limitations of Axiomatic Systems

We know a good deal now about symbolic languages and the logical procedures that they can express which was not known when I took the Mathematical Tripos at Cambridge in 1930. There were two of us who offered what was called Mathematical Philosophy that summer: Max Black and I. The man who had lectured to us had been the prodigious, prodigal Frank Ramsey. But he had died early that year, a month before his twenty-seventh birthday, and I imagine (though I cannot be sure) that we were examined in his place by his friend Richard Braithwaite. Whoever the examiner was, he blandly asked us on one of our papers to discuss the *Entscheidungsproblem.*

The *Entscheidungsproblem*, the problem of decision, was a startling question which David Hilbert had posed: whether it is self-evident—whether indeed it can be shown—that all mathematical assertions which make sense can necessarily be proved to be either true or false. The question had gone unanswered for a long time, and neither Max Black nor I was likely to settle it at short notice in an afternoon. I no longer remember what general arguments I produced in the examination room for and against the disputed possibility. For history caught up with us and our examiner in a spectacular way, ironically within a year.

Most professional scientists now know what happened. In 1931 a young Austrian mathematician, Kurt Gödel, proved two remarkable and remarkably unwelcome theorems. The first theorem says that any logical system which is not excessively simple (that is, which at least includes ordinary arith-

metic) can express true assertions which nevertheless cannot be deduced from its axioms. And the second theorem says that the axioms in such a system, with or without additional truths, cannot be shown in advance to be free from hidden contradictions. In short, a logical system which has any richness can never be complete, yet cannot be guaranteed to be consistent.

That was in 1931. In the next few years, other unpleasant theorems were established. A. M. Turing in England and Alonzo Church in America showed that no mechanical procedure can be devised that can test every assertion in a logical system and in a finite number of steps demonstrate it to be either true or false. This is Hilbert's *Entscheidungsproblem* in its direct form. In a sense, Gödel's result is deeper than this; and Alfred Tarski in Poland proved an even deeper limitation of logic. Tarski showed that there can be no precise language that is universal, every formal language which is at least as rich as arithmetic contains meaningful sentences that cannot be asserted to be either true or false.

In order to leave no room for doubt, let me linger on the essential content of these extraordinary and far-reaching theorems. They are theorems in mathematical logic, and in one sense the mathematics cannot be removed from them. That is to say, any logical system to which they apply must include the arithmetic of whole numbers as a basic and distinguishable part. But with this proviso, to which I shall return, they apply to any system of thought which attempts to set up a basis of fundamental axioms and then to match the world by making deductions from them in an exact language—the language of physics, for example, or the chemical language inside the brain.

Such a system of axioms has always been thought to be the ideal model for which all science strives. Indeed, it could be said that theoretical science is the attempt to uncover an ultimate and comprehensive set of axioms (including mathemati-

cal rules) from which all the phenomena of the world could be shown to follow by deductive steps. But the results that I have quoted, and specifically the theorems of Gödel and of Tarski, make it evident that this ideal is hopeless. For they show that every axiomatic system of any mathematical richness is subject to severe limitations, whose incidence cannot be foreseen and yet which cannot be circumvented. In the first place, not all sensible assertions in the language of the system can be deduced (or disproved) from the axioms: no set of axioms can be complete. And in the second place, an axiomatic system can never be guaranteed to be consistent: any day, some flagrant and irreconcilable contradiction may turn up in it. An axiomatic system cannot be made to generate a description of the world which matches it fully, point for point; at some points there will be holes that cannot be filled in by deduction, and at other points two opposite deductions may turn up.

Implications for Science

The implications of these results for any theory of knowledge have long been stressed, for example by Rudolf Carnap and by Karl Popper. But in addition I am stressing here, as I have done before (in *The Common Sense of Science* in 1951), their implications for empirical science. For I believe that any exact science must include in its system the axioms of arithmetic, in the form of procedures which require us to distinguish all the whole numbers. For example, if we seek to reduce all the sciences to physics, then we shall need the theory of groups and the statistics of assemblies of particles; and both these operations are subject to Gödel's theorems. In the same way, the statistical limitation on the recurrence of physical systems which Henri Poincaré first demonstrated in ergodic theory are, in my view, another expression of Turing's and Church's theorem that it is impossible to decide for every in-

stance whether it is a consequence of the axioms. And finally, Tarski's theorem demonstrates, I think conclusively, that there cannot be a universal description of nature in a single, closed, consistent language.

I hold, therefore, that the logical theorems reach decisively into the systemization of empirical science. It follows in my view that the unwritten aim that the physical sciences have set themselves since Isaac Newton's time cannot be attained. The laws of nature cannot be formulated as an axiomatic, deductive, formal, and unambiguous system which is also complete. And if at any stage in scientific discovery the laws of nature did seem to make a complete system, then we should have to conclude that we had not got them right. Nature cannot be represented in the form of what logicians now call a Turing machine—that is, a logical machine operating on a basic set of axioms by making formal deductions from them in an exact language. There is no perfect description conceivable, even in the abstract, in the form of an axiomatic and deductive system.

Of course, we suppose nevertheless that nature does obey a set of laws of her own which are precise, complete, and consistent. But if this is so, then their inner formulation must be of some kind quite different from any that we know, and at present we have no idea how to conceive it. Any description in our present formalisms must be incomplete, not because of the obduracy of nature, but because of the limitation of language as we use it. And this limitation lies not in the human fallibility of language but, on the contrary, in its logical insufficiency.

This is a cardinal point: it is the language that we use in describing nature that imposes (by its arrangement of definitions and axioms) both the form and the limitations of the laws that we find. For example, it may be held that if we can remove the arithmetic from physics, we may yet get an axiomatic system which is complete and consistent. I do not share this view,

cal rules) from which all the phenomena of the world could be shown to follow by deductive steps. But the results that I have quoted, and specifically the theorems of Gödel and of Tarski, make it evident that this ideal is hopeless. For they show that every axiomatic system of any mathematical richness is subject to severe limitations, whose incidence cannot be foreseen and yet which cannot be circumvented. In the first place, not all sensible assertions in the language of the system can be deduced (or disproved) from the axioms: no set of axioms can be complete. And in the second place, an axiomatic system can never be guaranteed to be consistent: any day, some flagrant and irreconcilable contradiction may turn up in it. An axiomatic system cannot be made to generate a description of the world which matches it fully, point for point; at some points there will be holes that cannot be filled in by deduction, and at other points two opposite deductions may turn up.

Implications for Science

The implications of these results for any theory of knowledge have long been stressed, for example by Rudolf Carnap and by Karl Popper. But in addition I am stressing here, as I have done before (in *The Common Sense of Science* in 1951), their implications for empirical science. For I believe that any exact science must include in its system the axioms of arithmetic, in the form of procedures which require us to distinguish all the whole numbers. For example, if we seek to reduce all the sciences to physics, then we shall need the theory of groups and the statistics of assemblies of particles; and both these operations are subject to Gödel's theorems. In the same way, the statistical limitation on the recurrence of physical systems which Henri Poincaré first demonstrated in ergodic theory are, in my view, another expression of Turing's and Church's theorem that it is impossible to decide for every in-

stance whether it is a consequence of the axioms And finally, Tarski's theorem demonstrates, I think conclusively, that there cannot be a universal description of nature in a single, closed, consistent language.

I hold, therefore, that the logical theorems reach decisively into the systemization of empirical science. It follows in my view that the unwritten aim that the physical sciences have set themselves since Isaac Newton's time cannot be attained. The laws of nature cannot be formulated as an axiomatic, deductive, formal, and unambiguous system which is also complete. And if at any stage in scientific discovery the laws of nature did seem to make a complete system, then we should have to conclude that we had not got them right. Nature cannot be represented in the form of what logicians now call a Turing machine—that is, a logical machine operating on a basic set of axioms by making formal deductions from them in an exact language There is no perfect description conceivable, even in the abstract, in the form of an axiomatic and deductive system

Of course, we suppose nevertheless that nature does obey a set of laws of her own which are precise, complete, and consistent. But if this is so, then their inner formulation must be of some kind quite different from any that we know, and at present we have no idea how to conceive it Any description in our present formalisms must be incomplete, not because of the obduracy of nature, but because of the limitation of language as we use it And this limitation lies not in the human fallibility of language but, on the contrary, in its logical insufficiency

This is a cardinal point it is the language that we use in describing nature that imposes (by its arrangement of definitions and axioms) both the form and the limitations of the laws that we find. For example, it may be held that if we can remove the arithmetic from physics, we may yet get an axiomatic system which is complete and consistent I do not share this view,

cal rules) from which all the phenomena of the world could be shown to follow by deductive steps. But the results that I have quoted, and specifically the theorems of Gödel and of Tarski, make it evident that this ideal is hopeless. For they show that every axiomatic system of any mathematical richness is subject to severe limitations, whose incidence cannot be foreseen and yet which cannot be circumvented. In the first place, not all sensible assertions in the language of the system can be deduced (or disproved) from the axioms: no set of axioms can be complete. And in the second place, an axiomatic system can never be guaranteed to be consistent: any day, some flagrant and irreconcilable contradiction may turn up in it. An axiomatic system cannot be made to generate a description of the world which matches it fully, point for point; at some points there will be holes that cannot be filled in by deduction, and at other points two opposite deductions may turn up.

Implications for Science

The implications of these results for any theory of knowledge have long been stressed, for example by Rudolf Carnap and by Karl Popper. But in addition I am stressing here, as I have done before (in *The Common Sense of Science* in 1951), their implications for empirical science. For I believe that any exact science must include in its system the axioms of arithmetic, in the form of procedures which require us to distinguish all the whole numbers. For example, if we seek to reduce all the sciences to physics, then we shall need the theory of groups and the statistics of assemblies of particles; and both these operations are subject to Gödel's theorems. In the same way, the statistical limitation on the recurrence of physical systems which Henri Poincaré first demonstrated in ergodic theory are, in my view, another expression of Turing's and Church's theorem that it is impossible to decide for every in-

stance whether it is a consequence of the axioms. And finally, Tarski's theorem demonstrates, I think conclusively, that there cannot be a universal description of nature in a single, closed, consistent language.

I hold, therefore, that the logical theorems reach decisively into the systemization of empirical science. It follows in my view that the unwritten aim that the physical sciences have set themselves since Isaac Newton's time cannot be attained. The laws of nature cannot be formulated as an axiomatic, deductive, formal, and unambiguous system which is also complete. And if at any stage in scientific discovery the laws of nature did seem to make a complete system, then we should have to conclude that we had not got them right. Nature cannot be represented in the form of what logicians now call a Turing machine – that is, a logical machine operating on a basic set of axioms by making formal deductions from them in an exact language. There is no perfect description conceivable, even in the abstract, in the form of an axiomatic and deductive system.

Of course, we suppose nevertheless that nature does obey a set of laws of her own which are precise, complete, and consistent. But if this is so, then their inner formulation must be of some kind quite different from any that we know, and at present we have no idea how to conceive it. Any description in our present formalisms must be incomplete, not because of the obduracy of nature, but because of the limitation of language as we use it. And this limitation lies not in the human fallibility of language but, on the contrary, in its logical insufficiency.

This is a cardinal point: it is the language that we use in describing nature that imposes (by its arrangement of definitions and axioms) both the form and the limitations of the laws that we find. For example, it may be held that if we can remove the arithmetic from physics, we may yet get an axiomatic system which is complete and consistent. I do not share this view,

cal rules) from which all the phenomena of the world could be shown to follow by deductive steps. But the results that I have quoted, and specifically the theorems of Gödel and of Tarski, make it evident that this ideal is hopeless. For they show that every axiomatic system of any mathematical richness is subject to severe limitations, whose incidence cannot be foreseen and yet which cannot be circumvented. In the first place, not all sensible assertions in the language of the system can be deduced (or disproved) from the axioms: no set of axioms can be complete. And in the second place, an axiomatic system can never be guaranteed to be consistent: any day, some flagrant and irreconcilable contradiction may turn up in it. An axiomatic system cannot be made to generate a description of the world which matches it fully, point for point; at some points there will be holes that cannot be filled in by deduction, and at other points two opposite deductions may turn up.

Implications for Science

The implications of these results for any theory of knowledge have long been stressed, for example by Rudolf Carnap and by Karl Popper. But in addition I am stressing here, as I have done before (in *The Common Sense of Science* in 1951), their implications for empirical science. For I believe that any exact science must include in its system the axioms of arithmetic, in the form of procedures which require us to distinguish all the whole numbers. For example, if we seek to reduce all the sciences to physics, then we shall need the theory of groups and the statistics of assemblies of particles; and both these operations are subject to Gödel's theorems. In the same way, the statistical limitation on the recurrence of physical systems which Henri Poincaré first demonstrated in ergodic theory are, in my view, another expression of Turing's and Church's theorem that it is impossible to decide for every in-

stance whether it is a consequence of the axioms. And finally, Tarski's theorem demonstrates, I think conclusively, that there cannot be a universal description of nature in a single, closed, consistent language.

I hold, therefore, that the logical theorems reach decisively into the systemization of empirical science. It follows in my view that the unwritten aim that the physical sciences have set themselves since Isaac Newton's time cannot be attained. The laws of nature cannot be formulated as an axiomatic, deductive, formal, and unambiguous system which is also complete. And if at any stage in scientific discovery the laws of nature did seem to make a complete system, then we should have to conclude that we had not got them right. Nature cannot be represented in the form of what logicians now call a Turing machine—that is, a logical machine operating on a basic set of axioms by making formal deductions from them in an exact language. There is no perfect description conceivable, even in the abstract, in the form of an axiomatic and deductive system.

Of course, we suppose nevertheless that nature does obey a set of laws of her own which are precise, complete, and consistent. But if this is so, then their inner formulation must be of some kind quite different from any that we know, and at present we have no idea how to conceive it. Any description in our present formalisms must be incomplete, not because of the obduracy of nature, but because of the limitation of language as we use it. And this limitation lies not in the human fallibility of language but, on the contrary, in its logical insufficiency.

This is a cardinal point. it is the language that we use in describing nature that imposes (by its arrangement of definitions and axioms) both the form and the limitations of the laws that we find. For example, it may be held that if we can remove the arithmetic from physics, we may yet get an axiomatic system which is complete and consistent. I do not share this view,

but it is arguable; yet it does not seem to me to bear in fact on our present formulation of the laws of nature. On present evidence, we must conclude (in my view) that the human mind is constrained to conceive physical laws in arithmetical language: the whole numbers are literally an integral part of its conceptual apparatus. If this is so, then the mind cannot extricate the laws of nature from its own language—our formal logic is not that of nature—and we are not at all, as Leibnitz and others have thought, in a "pre-established harmony" with the language of nature.

Adding a New Axiom

Every scientific system, as we understand that phrase now, is incomplete: simply as a logical machine, it cannot cover all the phenomena of nature. It therefore follows, not merely in practice but in principle, that the system must be enlarged, from time to time by the addition of new axioms, which cannot however be foreseen or proved to be free from contradictions. How does the outstanding scientist come to propose such a decisive axiom, while less imaginative minds go on tinkering with the old system? How did Gregor Mendel leap to conceive the statistical axioms of genetics? What moved Albert Einstein to make the constancy of the speed of light not a consequence but an axiom in the construction of relativity?

An obvious answer is that the great mind, like the small, experiments with different alternatives, works out their consequences for some distance, and thereupon guesses (much like a chess player) that one move will generate richer possibilities than the others. But this answer only shifts the question from one foot to the other. It still remains to ask how the great mind comes to guess better than another, and to make leaps that turn out to lead farther and deeper than yours or mine.

We do not know; and there is no logical way we can know,

or can formalize the pregnant decision. The step by which a new axiom is added cannot itself be mechanized. It is a free play of the mind, an invention outside the logical processes. This is the central act of imagination in science, and it is in all respects like any similar act in literature; it can in fact be taken as a definition of imagination. In this respect, science and literature are alike: in both of them, the mind decides to enrich the system as it stands by an addition which is made by an unmechanical act of free choice.

As for the invention that is added—the new relation in science or the imaginative shift of vision in literature—its birth is always the same. It begins in the multiple meanings and overtones, the hidden ambiguities, which human language contains in spite of our best efforts to make it sharp. The language of thought consists for the most part of general words, and although such a word may be as matter of fact as *parallel* or as solid as *mass*, as down to earth as *table*, there is always about it a penumbra of uncertainty and ambivalence from which new relations may suddenly become apparent. *Parallel* may become the beginning for non-Euclidean geometries, and *mass* may become equivalent to energy, for the universal reason that even a *table* cannot be defined in terms which allow us to say with absolute decision of every object in the universe that it is either a table or not a table. Frank Ramsey, of whom I spoke earlier, proved that this is an indispensable factor in the development of any science, and in this important sense, he anticipated some of the implications of Godel.

It is characteristic of human language that it is made up of past metaphors and analogies, and they are a fertile ground for the exploration of ambiguity and the discovery of hidden likenesses. Here begin the unexpected links and conjunctions which literature (and all art) constantly produces, and the inventive ideas of science begin here too.

How these ambivalences are developed in science and in

but it is arguable; yet it does not seem to me to bear in fact on our present formulation of the laws of nature. On present evidence, we must conclude (in my view) that the human mind is constrained to conceive physical laws in arithmetical language: the whole numbers are literally an integral part of its conceptual apparatus. If this is so, then the mind cannot extricate the laws of nature from its own language—our formal logic is not that of nature—and we are not at all, as Leibnitz and others have thought, in a "pre-established harmony" with the language of nature.

Adding a New Axiom

Every scientific system, as we understand that phrase now, is incomplete: simply as a logical machine, it cannot cover all the phenomena of nature. It therefore follows, not merely in practice but in principle, that the system must be enlarged from time to time by the addition of new axioms, which cannot however be foreseen or proved to be free from contradictions. How does the outstanding scientist come to propose such a decisive axiom, while less imaginative minds go on tinkering with the old system? How did Gregor Mendel leap to conceive the statistical axioms of genetics? What moved Albert Einstein to make the constancy of the speed of light not a consequence but an axiom in the construction of relativity?

An obvious answer is that the great mind, like the small, experiments with different alternatives, works out their consequences for some distance, and thereupon guesses (much like a chess player) that one move will generate richer possibilities than the others. But this answer only shifts the question from one foot to the other. It still remains to ask how the great mind comes to guess better than another, and to make leaps that turn out to lead farther and deeper than yours or mine.

We do not know; and there is no logical way we can know,

or can formalize the pregnant decision. The step by which a new axiom is added cannot itself be mechanized. It is a free play of the mind, an invention outside the logical processes. This is the central act of imagination in science, and it is in all respects like any similar act in literature; it can in fact be taken as a definition of imagination. In this respect, science and literature are alike: in both of them, the mind decides to enrich the system as it stands by an addition which is made by an unmechanical act of free choice.

As for the invention that is added—the new relation in science or the imaginative shift of vision in literature—its birth is always the same. It begins in the multiple meanings and overtones, the hidden ambiguities, which human language contains in spite of our best efforts to make it sharp. The language of thought consists for the most part of general words, and although such a word may be as matter of fact as *parallel* or as solid as *mass*, as down to earth as *table*, there is always about it a penumbra of uncertainty and ambivalence from which new relations may suddenly become apparent. *Parallel* may become the beginning for non-Euclidean geometries, and *mass* may become equivalent to energy, for the universal reason that even a *table* cannot be defined in terms which allow us to say with absolute decision of every object in the universe that it is either a table or not a table. Frank Ramsey, of whom I spoke earlier, proved that this is an indispensable factor in the development of any science, and in this important sense, he anticipated some of the implications of Gödel.

It is characteristic of human language that it is made up of past metaphors and analogies, and they are a fertile ground for the exploration of ambiguity and the discovery of hidden likenesses. Here begin the unexpected links and conjunctions which literature (and all art) constantly produces, and the inventive ideas of science begin here too.

How these ambivalences are developed in science and in

literature is the theme of *The Identity of Man*, and I can only summarize it here. In science, the aim is to disentangle each ambiguity, and to force nature to decide between the alternatives by a critical experiment. In this way, we progress in science (as it were) by turning the information from nature through the logical machine of the brain into an effective tape instruction. In literature the ambiguities are not resolved, and the brain works or plays with the information without ever turning it into a machine instruction. But in both, the new invention is taken by the same kind of step, and at the moment when the step is taken we are in no logical system: we have left one system and are about to enter and form the other, and are in a no-man's-land outside logic.

Paradoxes of Self-Reference

The first half of my theme, which I have now completed, has consisted of theorems in mathematical logic and their application first to the language of science and then, incidentally, to literature. What I have shown there, the surprising demand that they imply for a kindred imagination in both, is unsettling, of course, and awkward, because this is not at all how we wanted the grand panorama of knowledge to look. But there it is, we must come to terms with it; and so far I have simply displayed what the terms are, as a matter of fact.

Now I turn to the second part of my theme, to discuss a sharply different aspect of the same problem. I shall still be concerned with these maverick theorems in logic, but with something else about them: not so much with their existence and implications as with their origin. For there is a common source from which all these theorems spring, and it is uncommonly interesting and revealing.

Specifically, the two theorems of Gödel, the theorems of Turing and Church, and Tarski's theorem say different

things. Each of them establishes some limitation on a logical system, either on its completeness or its consistency, and these limitations are not quite the same. Yet they do form a common family of limitations, and this is because they all arise from a common difficulty in all symbolic language. The difficulty is that the language can be used to describe not only parts of the world but also parts of the language itself. In each of them, the proof hinges on a construction by which a proposition *about* arithmetic is expressed as a proposition *in* arithmetic.

Many logical problems grow from this common root, namely that the range of reference of any reasonably rich system necessarily includes reference to itself. This creates an endless regress, an infinite hall of mirrors of self-reflection. And the regress comes sharply to a focus in all the paradoxes of logic, which are cousins of one sort or another to the classical contradiction that the Greeks knew what they called the Cretan paradox. This is the contradiction implied by the statement of Epimenides the Cretan that all Cretans are liars.

There are many modern forms of this and its related paradoxes. One form is Bertrand Russell's definition of the class of all classes that are not members of themselves. Another is the paradox of Jules Richard, which (roughly) gives this a numerical dress. Godel constructed his theorems on this pattern. Perhaps the punning, linguistic quality of these contradictions, their oddly literary playfulness, is best displayed by a remark in the same vein by Groucho Marx, who said that he would not think of belonging to a club that was willing to have him for a member. Yet these are not trivial matters: they face us whenever we contrast rules and exceptions, tolerance and intolerance, and all the human issues which join and divide us in argument at the same time.

The mathematical paradoxes, and the devices derived from them that Godel and others exploited for their theorems, all have the same feature: they depend on the use of

concepts whose range of reference includes the concept itself. In short, the model for them all is the Cretan paradox, the simple sentence, "What I am now saying is not true." This is obviously a self-contradiction: if the assertion is true, then by its own evidence it is not true; and if the assertion is false, then that tells us that what is being said must be true.

Bertrand Russell tried (with Alfred North Whitehead in the *Principia Mathematica*) to untie the knot in this kind of paradox, and to put an end to the infinite regress of assertions about assertions, by constructing a theory of types. This was intended to prevent us from using the same language to discuss our language that we use to discuss the things that the language names. A hierarchy of types was created, starting with simple sentences about things, going on to sentences about sentences about things, then to sentences about sentences which are themselves about sentences about things, and so on. No one could look on this infinite construction with anything but a suspicious eye, and so it turned out; the theory of types is an unhappy artifice. If as human beings we want to use human language, then we must accept that part of its richness is in its capacity to refer to itself.

I stress, in what I have just said, the word *human.* Animals use language to signal to one another, and what they have to say essentially refers to states of affairs (factual or emotional) and to nothing else. Such a language has no problems of self-reference: it is intended to pass information from one animal to another, directly and unequivocally as an instruction. In this sense, René Descartes was right to say that animals are machines and human beings are not. Human language is richer precisely because we think about ourselves. We cannot eliminate self-reference from human language without thereby turning it from a genuine language of information into a machine language of instructions.

In particular, all philosophy and epistemology operates by its nature within the field where the difficulties lie, the field

of self-reference. I mean by self-reference the construction of sentences, in thought or in speech, whose range of application includes that very kind of sentence. On this definition, "I am hungry" contains no self-reference, but "I am troubled" does. All thinking about thinking implies self-reference: the first statement of principle in the philosophy of Descartes, *Cogito ergo sum*, refers to itself. It is this very cogitation, or the class of cogitations that includes it, which gives the speaker the right to assert that he is cogitating. Philosophy is not possible without the regress of cogitation about cogitation. Whatever could be thought by machines, philosophy certainly could not. Indeed, on my view of human language, philosophy could not even be thought about by animals.

Philosophy and Psychology

It is clear enough that statements in philosophy are, by their nature, often dogged by self-reference, and that philosophy as a discipline is therefore limited even more severely than science by the logical gaps that the theorems of Godel and Tarski have laid bare. In mathematics and science, it is a surprise to find oneself bounded by these theorems; it is not at all obvious, and indeed is unexpected, to learn that mathematical and scientific statements cannot be wholly cleared of self-references (or of some equivalent recursive regress). But it is evident from the outset that philosophy is full of self-references, and therefore that, if the breakdown in the machinery of logic has its origin in self-reference, philosophy is surely subject to it. Indeed it is clear that while mathematics and science are subject to it only from time to time, when a new step has to be taken, philosophy is subject to it severely and constantly—because self-reference is built into its very method.

In the same way, we can see at once why psychology and psychoanalysis, regarded as sciences, are most severely sub-

ject to the theorems of logical limitation. There was a time when no clear boundary was drawn between philosophy and psychology; Thomas Hobbes, John Locke, and David Hume all wrote philosophy, much of which was a study of the mind, and was, for its age, a form of psychology. Now that psychology has entered into less conscious fields of the mind, the logical problems that are created by self-reference are very patent. Many natural scientists complain that psychology, and other studies of human thought and behavior, lack the rigor of a true science. This is usually excused on the ground that such human studies are young, and have not yet developed the proper formal apparatus by which information can be turned into exact prediction. But I suggest that the logical theorems now show us that this excuse is mistaken. There is an essential difficulty in casting these disciplines into an axiomatic system; they are limited, more severely and more constantly than the natural sciences, by the self-reference that underlies them everywhere. And it cannot be got out of the system by the occasional addition of a new axiom, as in the natural sciences. The axiomatic method as such may be unworkable in these studies, and whatever machinery is discovered for them in the future will (I think) not be of this traditional kind.

This is illustrated by Karl Popper's account of how he became disillusioned with the psychoanalytic explanations of Sigmund Freud and Alfred Adler. In natural science, remarked Popper, a theory is expected to make a prediction, and one prediction only, about the outcome of an experiment; and it is discarded if this forecast is not fulfilled in the experiment. But the theories of psychoanalysis are not of this kind at all; as Popper found, they are constantly explaining that my neighbor on the right is polite because he has an inferiority complex, and my neighbor on the left is rude because he has an inferiority complex. If therefore I turn the concept of the inferiority complex around, I get the unhelp-

ful prediction that it may cause my neighbors either to be polite or to be rude. This is not what we expect of a scientific theory. And indeed it is not: all arguments derived from Freud's invention of the unconscious have this paradoxical content, precisely because their use of self-reference creates paradoxes. The Cretan who said that all Cretans are liars was talking a classical form of the language in which the psychoanalyst frames such concepts as the unconscious and its inferiority complex. If he had lived not in Crete but in Vienna, he would have said that all Viennese have inferiority complexes.

Self-Reference in Literature

Beyond these borderline fields stands, full face, the particular interest to which I am drawn: the art of literature. A work of literature is in the first place a description or a story. William Wordsworth's poem *The Daffodils* is a description, and the *Oedipus Rex* of Sophocles is a story. Neither a description nor a story need contain any overt self-references. For example, a description of my interests and the story of my career are neutral accounts which do not demand that you involve yourself by referring any part of them to yourself. Unhappily, when that has been said about this description and this story, nothing at all has been said either about *The Daffodils* or about *Oedipus Rex*. Yes, it is possible to have descriptions and stories whose content does not draw us into them, and in which our minds do not reflect on themselves, but such accounts simply do not have the power of Wordsworth and Sophocles. Neither, I am afraid, will they have their immortality.

From these simple examples it is at once clear that literature is literature only when it demands and commands our personal involvement. It is insistent because it insists that the bland descriptions of flowers and the remote *Police Gazette* records of incest and suicide concern us. They become part of us and we of them, they draw us to the human race and the

human condition, and they make us one with Wordsworth on his couch and Jocasta in her bed and the plague-racked men of Thebes all over the world.

What is true of literature is true of every art. The work of art is a constructed thing, and is so even when it happens to have been found in nature in the form into which we now read a human meaning. It has been made in essence, its meaning has been created, by a human being: it expresses his vision of the relation between man and nature, and it invites us not to like it or to dislike it but to be drawn into it. The work of art compels us (when it is compelling) to look at the world with it, and to look through it into the mind of its maker. We cannot dissociate the work from its origin, which is to be a made thing—a thing made by a man which expresses how the man sees himself in the world. It interests us only as it engages us, and asks us to see ourselves in the same world also. Although what is expressed in the work is another man's self, the reference is to ourselves because the reference is universally to the human self.

Let me be explicit in my meaning here. I am not merely remarking that there is self-reference in the moral reflections of the Greek chorus, or in the reflections, "In vacant or in pensive mood" that fill Wordsworth's inward eye in solitude. These are only of the same kind as the reflections of Descartes in philosophy and of an analyst interpreting dreams. But the self-references of literature, and of art in general, go deeper than these formal thoughts. My argument is that literature is composed essentially of self-reference, and takes its life from the dual tension between watching our own minds from the inside and watching someone else's from the outside. And this is one of the classical paradoxes in the theory of knowledge, how and when we know that others do indeed feel as we do, which Ludwig Wittgenstein for example discussed in the *Blue Book* long before I discussed it in *The Identity of Man*.

The force and meaning of literature is to present the lives of others to us in such a way that we recognize ourselves in them, and live them from the outside and from the inside together. We do not understand Wordsworth unless our heart also turns over at the golden host, and the tragedy of Oedipus differs from the gunplay in the Sunday paper only if we recognize ourselves in the characters. We have to see that Oedipus is us, capable of killing a stranger at the crossroads and blundering into a labyrinth of horror. We have to see that Jocasta is us, longing for the lost youth who so transparently is a part of herself in both senses: the son, who is also the symbol of her own youth, that she longs to recapture and sense again in her leaping womb. And when we recognize this in Jocasta and in ourselves, it is more tender, more heartbreaking, more deeply human than the explanations of psychoanalysis. Of course Freud was right about the Oedipus complex, but Sophocles wakes deeper echoes than Freud, because he brings home to us the longing of Jocasta for herself—the self that she was and the self that she gave birth to—in the same hushed breath with the familiar and family jealousy of Oedipus.

Literature and art live by, they come alive in, the sense of our own self stretching into the actions and disasters of someone else's self, and thereby mapping the human self as a whole. This is how I put this part of my theme in *The Identity of Man*:

> I hold that each man has a self, and enlarges his self by his experiences. That is, he learns from experience: from the experiences of others as well as his own, and from their inner experiences as well as their outer. But he can learn from their inner experience only by entering it, and that is not done merely by reading a written record of it. We must have the gift to identify ourselves with other men, to relive their experience and to feel its conflicts as our own. And the conflicts are the essence of

the experience. We gain knowledge of ourselves by identifying ourselves with others, but that is not enough—that only gives us the fantasies of sex and the parodies of power, the absurd strutting daydreams of Secret Agent 007 and *Butterfield 8*. We must enter others in order to share their conflicts, and they must be shown to have grave conflicts, in order that we shall feel in their lives what we know in our own: the human dilemma. The knowledge of self cannot be formalized because it cannot be closed, even provisionally; it is perpetually open, because the dilemma is perpetually unresolved.

Science and Literature

Let me recapitulate the steps in my argument. I treated my theme in two parts: the first was concerned in the main with science, and the second with literature. In both parts I was at pains to show that the brain as a machine is certainly not the kind of machine that we understand now. It is not a logical machine, because no logical machine can reach out of the difficulties and paradoxes created by self-reference. The logic of the mind differs from formal logic in its ability to overcome and indeed to exploit the ambivalences of self-reference, so that they become the instruments of imagination.

In the first half of my theme, I explained the limitations (they derive from self-reference) that circumscribe any axiomatic and deductive system of a reasonable richness, in mathematics and (I hold) in the natural sciences. The logical theorems which I quoted and explained show that this must be so, and they also show how these logical gaps have to be filled, and new theorems incorporated as added axioms in a system, at each step. The decision to take new matter into our systems, in science or in literature, has no analog in any logical machine. It is an imaginative step of a kind that we do not understand but that we can watch in the work of a great scientist

for a great writer; and it is alike in science and in literature.

The second part of my theme goes farther. Here I pointed out that human language, when it is specifically human and is concerned with reflection and judgment about our own lives, is necessarily full of self-references. This is clear in philosophy and in psychology. But it reaches deeper in literature, because the essence of literature (and of all art) lies in the identification of ourselves with other human beings whose actions we are watching and judging as if they were our own. Here the self-reference is so integral that we cannot construct any of the provisional systems with which mathematics and science make do for a time, and which they then amend when the need arises.

In literature, there is no provisional description that can take the place of the work itself. We cannot replace it, as in science, by an axiomatic system which will do until it turns out to fall short and has to be enlarged. The references in literature by the writer to himself and others, and by the reader to himself in what he reads, penetrate the work through and through; and there is no way of getting round Godel's theorems and Tarski's theorem and the others by any step-by-step procedure. In this respect, science and literature are different.

Neither science nor literature ever gives a complete account of nature or of life. In both of them, the progress from the present account to the next account is made by the exploration of the ambiguities in the language that we use at this moment. In science, these ambiguities are resolved for the time being, and a system without ambiguity is built up provisionally, until it is shown to fall short. This is why the results of science at any given moment can be presented on an axiomatic and deductive machine, although nature as a whole can never be so presented because no such machine can be complete. Whatever kind of machine nature is, it is different from this.

But in literature, the ambiguities cannot be resolved even

for the time being, and no provisional system of axioms can be set up to describe the human situation as the writer and the reader seek to see it together. Here the brain cannot act as a logical machine even for the time being, by which I mean that it cannot take in the information, sort out its ambiguities, and turn it into unambiguous instructions. That is not what a work of art does to us, and we cannot derive such instructions from it. I will quote at the last the passage from *The Identity of Man* of which, as I promised at the beginning, this essay is a detailed exposition. It states for the machinery of the brain the same limitations that I have exhibited in its account of the machinery of nature.

> I am asserting that there is a mode of knowledge which cannot be spelled out formally to direct a machine. It may be asked, Any machine? If this is a question in the present, then the answer is Yes. For example, we know (from the work of Kurt Gödel and A. M. Turing) that no machine that uses strict logic can examine its own instructions and prove them consistent. But is it is a question about the measureless future, then it cannot be answered. A machine is not a natural object; it is a human artifact which mimics and exploits our own understanding of nature; and we cannot foresee how radically we may come to change that understanding. We cannot foresee and we cannot conceive all possible machines—if indeed the word *all* has a meaning in this sentence. All that we can say, and all that I can assert, is that we cannot now conceive any kind of law or machine which could formalize the total modes of human knowledge.

There is however one respect in which my exposition now goes radically beyond this passage, not merely in detail but in substance. That is in tracing the common quality of imagination in science and in literature to the logic of self-refer-

ence, and in showing that, within this common quality, the difference of mode between science and literature reflects the different extent to which self-reference enters their languages.

REFERENCES

1. R. B. Braithwaite, *Scientific Explanation*, Cambridge (1953)
2. J. Bronowski, *The Identity of Man*, New York (1965)
3. Rudolf Carnap, *The Logical Syntax of Language*, New York (1937)
4. Alonzo Church, A Note on the Entscheidungsproblem, *Symbolic Logic*, *1*, 40, 101 (1936)
5. Kurt Gödel, Ueber formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I, *Monatsh. Mathem. Physik*, *38*, 173 (1931)
6. David Hilbert and P. Bernays, *Grundlagen der Mathematik*, Berlin (1934-39)
7. S. C. Kleene, Recursive Predicates and Quantifiers, *Trans. Am. Math. Soc.*, *53*, 41 (1943)
8. John Myhill, Some Philosophical Implications of Mathematical Logic, *Rev. Metaphys.*, *6*, 165 (1952)
9. Ernest Nagel and James R. Newman, *Gödel's Proof*, New York (1958)
10. Henri Poincaré, *Calcul des Probabilités*, Paris (1912)
11. Karl R. Popper, *Conjectures and Refutations*, London (1963)
12. F. P. Ramsey, *The Foundations of Mathematics*, London (1931)
13. Jules Richard, Les principes des mathématiques et le problème des ensembles, *Rev. gén. sciences pures et appliquées*, *16*, 541 (1905)
14. Frédéric Riesz, Sur la théorie ergodique, *Commentarii Mathematici Helvetici*, *17*, 221 (1944-45)
15. Alfred Tarski, Der Wahrheitsbegriff in den formalisierten Sprachen, *Studia Philosophica*, *1*, 261 (1936)
16. A. M. Turing, On Computable Numbers with an Application to the Entscheidungsproblem, *Proc. London Math. Soc.*, Ser. 2, *42*, 230 (1936), *43*, 544 (1937)
17. A. N. Whitehead and Bertrand Russell, *Principia Mathematica*, Cambridge (1910-13)
18. Ludwig Wittgenstein, *The Blue and Brown Books*, Oxford (1958)

9. FROM GENE TO CHARACTER IN HIGHER PLANTS

By G. Ledyard Stebbins
University of California at Davis

Men have always been interested in finding out why a particular animal or plant looks the way it does. Biologists and laymen alike often try to answer such a question as "Why does the giraffe have such a long neck?" The answer a biologist would give to this question would depend upon his training and interests. A field naturalist might reply "Because this fits the giraffe to his environment by enabling him to get food from the leaves of tall trees." A Darwinian evolutionist would say "Because, in the remote past, certain animals with longer necks were better able to reach food, and so produced more offspring, to which they transmitted their longer necks." The answer of a developmental biologist might be "Because in the embryonic and fetal development a large quantity of growth hormones becomes concentrated in the neck region, thus stimulating the excessive development of the seven neck vertebrae and the tissues associated with them." Finally, a modern molecular geneticist would be likely to answer "Because a part of the code in the DNA of the giraffe's nuclei carries information for a long neck."

When we look over these answers one after the other, we can easily see that none of them is actually wrong, and none of them is more basic or important than the others. Each is essentially correct but incomplete. Surprisingly enough, how-

ever, one can easily find sophisticated, high-level discussions of similar biological problems in which one of these types of answers is considered to be the only significant biological one, and the others are minimized or completely neglected. Even when the need is recognized for these different approaches to such basic problems of biology, each type of answer is often regarded as sufficient in itself and separate from the others. The more we learn about the nature of life, however, the more we realize that these different avenues of approach are intimately connected with each other, and that each of the types of answers given above is incomplete unless it takes into account the facts which have led to the other answers. This intimate connection among form, function, development, heredity, and evolution is the main theme of the present discussion.

At the level of the macromolecule, enzymes provide striking examples of an elaborate, intricate, and highly specific structure associated with a precise and equally specific function. These molecules are the most directly connected with gene action of any of the working parts of the cell. They consist of one to four long chains of amino acid residues, the order of which in the chain is "coded" by the DNA of the genes, through the intermediary action of messenger RNA. This "primary structure" of the polypeptide chain is the ultimate basis of the very specific action of enzymes, but they achieve this specificity through an intricate series of relationships between different parts of the enzyme molecule and its substrate, which are only beginning to be understood. My account is based largely upon a recent review article by Koshland.[11]

The principal facts can be illustrated by the molecule of ribonuclease (Fig. 37), one of the best analyzed of enzyme molecules. This single polypeptide chain of 124 amino acid residues is completely inactive when stretched out. To be active, it must be folded into a very definite pattern, which

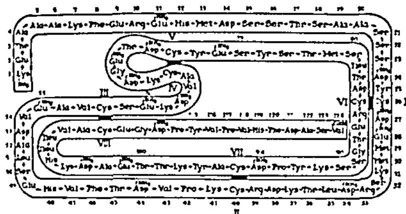


FIG. 37 The amino acid sequence in the polypeptide chain of ribonuclease [Figs 37 and 39 show the relationship between structure and function of enzyme molecules From D E Koshland in *Science, 142*, 1963 Reproduced by permission]

is determined by the fact that certain residues of cysteine in different parts of the molecule form disulfide (S–S) linkages with each other All polypeptide chains of active enzyme molecules must possess such a folded, secondary structure Molecules consisting of more than one polypeptide chain possess a tertiary structure because of similar linkages between particular residues on different chains

From this information we learn that certain amino acid residues which play no role at all in the reaction between enzyme and substrate are nevertheless of prime importance in determining the configuration of the enzyme molecule Their role, furthermore, depends upon their exact position in the polypeptide chain This is because the ability of the enzyme to become effectively attached to the molecules of the substrate upon which it works depends in a very precise fashion upon the configuration of the molecule The activity of the ribonuclease molecule in depolymerizing or breaking up the molecules of RNA is carried out by certain of its residues, which constitute the "active site" These are principally the

histidine at positions 12 and 119, the lysine at 41, the glutamine at 11, methionine at 13, aspartic acid at 121, and the alanine at 122. Some evidence suggests that the active enzyme is folded in such a way that positions 11 to 13 and 119 to 122 are rather close to each other and to the substrate. On the other hand, removal of all residues from positions 14 to 20 gives a molecule retaining 70 per cent of normal activity. Other enzyme molecules, such as that of papain, can suffer a loss of up to 70 per cent of their residues without seriously impairing their activity.

Experiments with enzyme molecules in which single amino acid residues at the active site have been chemically altered in various ways have shown that these residues may possess either one or two kinds of activity. Some of them hold the substrate in place, probably by forming covalent bonds with the substrate molecule, while others act directly upon that particular bond of the substrate molecule which the enzyme breaks or joins—in the case of ribonuclease, the bonds between nucleotides of RNA. Thus, the specificity of the enzyme molecule depends fundamentally upon three features of its gene-determined primary structure, or sequence of amino acid residues: (*1*) the position of those residues which, by forming linkages with each other, determine the secondary and tertiary structure of the molecule; (*2*) the position of those residues which form bonds with the substrate and must occupy a position on the folded chain that enables them to become attached to the substrate molecule; and (*3*) the position of those residues which interact chemically with the substrate molecule, and which, when the molecule is held in place, must be near to the part of it which they attack (Fig.38).

The most critical feature of the specificity of the enzyme molecule is the shape of the folded portion containing the active site. As Figure 39 shows, it must fit the substrate molecule in such an exact fashion that its active residues are properly placed in relation to both the substrate molecule and

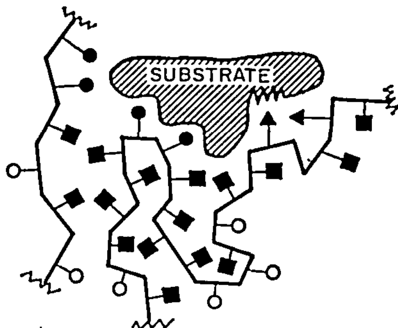


FIG. 38 Schematic representation of an active site. Solid circles are "contact" amino acids whose fit with substrate determines specificity, triangles are catalytic residues acting on substrate bond, indicated by a jagged line, open circles are nonessential residues on the surface, squares are residues whose interaction with each other maintains the three-dimensional structure of the protein. (From *J. Biol. Chem.*, reproduced by D. E. Koshland with the permission of Drs. Hirs, Stein, and Moore.)

to each other. For many years, the literature on enzymology has expressed this relationship by means of an analogy with a lock and key. Koshland, however, has pointed out the weakness of this analogy, since locks and keys are rigid structures whereas the polypeptide chains of enzyme molecules are highly flexible. A better analogy[1] is between the enzyme molecule and one of the highly complex tools used in modern surgery, such as the stapler for binding together two parts

1. For this analogy I am indebted to my colleague Prof. A. S. Fraser and to Dr. A. E. Hertle of the School of Veterinary Medicine, University of California, Davis.

of a blood vessel.[2] These tools are designed to fit tightly the blood vessel on which the operation is performed, and at the same time to insert a small staple for coupling the parts of the

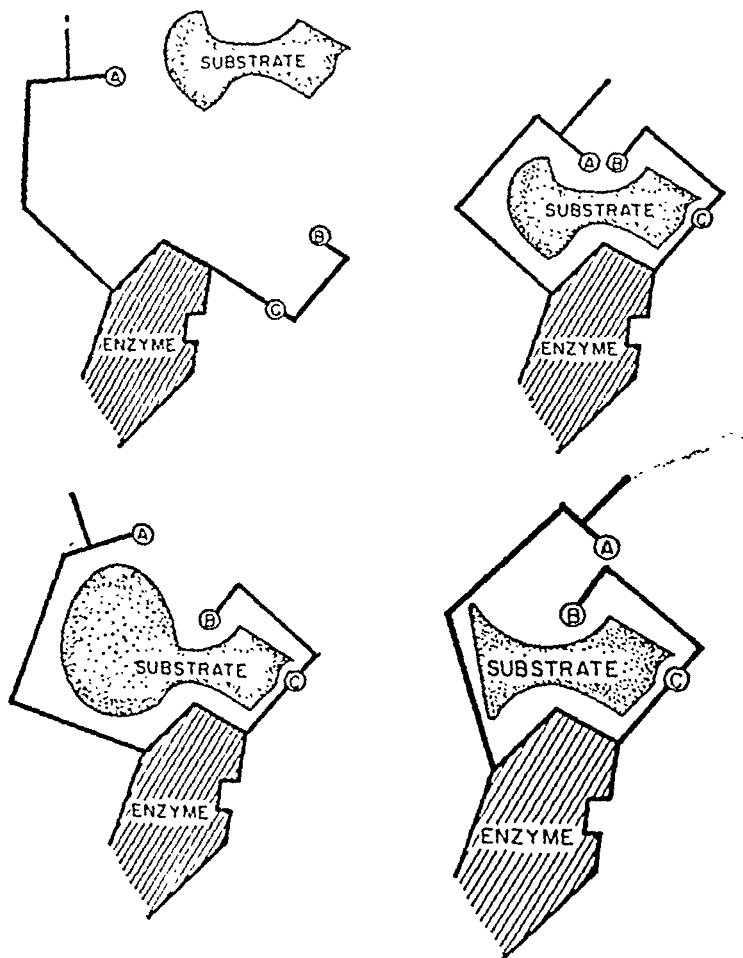


FIG. 39. Schematic representation of an active site. Substrate binding induces proper alignment of catalytic groups A and B so that reaction ensues (top). Compounds which are either too large or too small are bound but fail to cause proper alignment of catalytic groups, hence fail to react (bottom).

blood vessel together. The latter action is analogous to the formation of a chemical bond by an enzyme molecule with synthetic activity.

This analogy can be extended even further on the basis of the evidence analyzed by Monod et al.,[3] which indicates that the terminal product of a sequence of enzymatic reactions inhibits the enzyme responsible for the first reaction in the sequence, although it does not interact with the substrate on which the enzyme is working. Instead, this terminal metabolite apparently becomes bound to a part of the "inactive site" of the enzyme molecule or "allosteric protein," and by doing so alters the configuration of the entire molecule in such a way that it can no longer be active. If their allosteric protein model of feedback inhibition of enzyme activity proves to be generally applicable, then we can think of the inactive site of the enzyme molecule as analogous to the handle of the tool, and the terminal inhibiting metabolite as an automatic hand which releases the enzyme from its substrate when its work is done. The gene-determined order of amino acid residues thus appears to affect specificity in a fourth way. By determining the position, and perhaps the conditions under which a form-changing inhibitor can be bound to the otherwise inactive site of the polypeptide chain, it may regulate the activity of the enzyme in relation to that of other enzymatic reactions taking place in the cell.

The interrelationship of three of the four types of answers which might be given to the question "Why does the molecule of ribonuclease (or any other enzyme) have its particular *type of configuration?" is reasonably clear. The pattern of* folding serves the function of binding the molecule to its substrate and placing the active amino acid residues in the most advantageous position possible for their action on the particular chemical bond which is to be broken or joined. The "inactive sites" of the molecule may be important for the attachment of regulator or inhibitor molecules. Develop-

mentally, the pattern of folding is brought about by the attraction of certain amino acid residues for each other, thus forming the secondary and tertiary bonds. The hereditary basis for both these properties is the specific sequence of amino acid residues, which is determined by the DNA coding of the gene or genes which form the templates for the polypeptide chains.

The evolutionary answer to the specific structure of enzyme molecules is somewhat less clear, since it is hard to see how the structure of an enzyme molecule could change in such a way that it could become adapted to a new function without passing through several functionless intermediate structural arrangements. A possible answer was given many years ago by Weir,[4] who suggested that duplication of genes in the chormosomal complement could temporarily preserve functionless intermediate states at one of the duplicated loci, from which occasionally new gene structures could arise that would code for new enzymatic functions. This type of evolution, however, cannot have occurred very often. Most of the enzymatic reactions which the cells of higher animals or plants perform are also carried out by the cells of some species of bacteria and so were probably "invented" even before multicellular organisms appeared. Most probably, the evolution of enzyme molecules has involved chiefly their progressive modification to perform the same function in different cellular environments or in harmony with new combinations of other enzymes. This type of enzyme evolution has probably taken place through the successive appearance by genetic mutation of new isoenzymes. The discovery that "families" of isoenzymes may exist in a single organism, and that differentiation in higher organisms may be accompanied by the appearance of specific combinations of such enzymes,[5] is of the greatest significance for explaining the evolution of enzyme function. The most common type of enzyme evolution has probably been the replacement of one isoenzyme

bine, and one which links columbines to the related genera *Isopyrum* and *Paraquilegia*, is *A. ecalcarata*, a species found in eastern Asia, of which the petals completely lack spurs. In crosses between *A. ecalcarata* and either *A. vulgaris* or *A. canadensis*, the presence or absence of petal spur is inherited as a simple Mendelian character, so that the presence of the spur is governed by a singled dominant gene.[6] Furthermore, the difference between the bent spur of the European and the straight spur of the American species is also governed by a single gene (W. Gajewski, oral communication).

There are several reasons for believing that columbines originated in the Old World. Consequently, the first step in the evolution of the columbine flower from the white, anemone-like type of flower found in the related genera of the buttercup family may well have been the acquisition of purple color and the nodding position of the flower, as in the spurless Asiatic columbine. This was a partial adaptation to pollination by bumblebees, and was supplemented by the appearance and establishment of a dominant mutation for bent spurs on the petals. Columbines of this type then became widespread throughout Eurasia, and one of them migrated to North America. A purple-flowered columbine with short, bent spurs *(A. scopulorum)* still occurs in the Rocky Mountains. In the original North American populations, mutations for red color and straight spurs took place. When these became combined in populations growing in a region where bumblebees were scarce and hummingbirds plentiful, the action of natural selection established populations of columbines bearing flowers adapted to hummingbird pollination. Still another type of columbine flower, with pale color and very long spurs, is characteristic of several species found in our western mountains. These, the most specialized and probably most recent species of the genus, (e.g. *A. coerulea*, *A. chrysantha*, and *A. pubescens*), are adapted to pollination by nightflying hawk moths *(Sphingidae)*. The evolution of the columbine flower is

therefore closely associated with mutations affecting its color and structure, with migration of populations into new areas, and with natural selection for adaptation to new pollinators.

The developmental explanation of the differentiation of the flower, and its connection with gene action, are the least understood of any aspects of our problem. Various approaches to it will be discussed in the remainder of this article. The fact that strikes us most forcefully is that many of the genes responsible for differentiation produce their visible action only in very restricted parts of the plant. For instance, the genes for petal spurs in the columbine stimulate cell division in the petals but not in the sepals, even though sepals are produced almost at the same time as petals and are much more like them in other features of their adult structure. The genes for flower color in the red American columbines act in the upper part of the petal to produce yellow, and in the spur, the lower part of the same organ, to produce red color. Furthermore, localized differences in cells can be seen under the microscope in the form of large hairs which are scattered over the surface of the flowers. This means that, at a particular stage in development, certain cells receive genetic information that "tells" them to enlarge greatly and to assume a very distinctive shape, while their neighbors enlarge much more slowly and keep approximately the same shape they had when formed.

Gene Action and Differentiation

These facts raise for us a question which has been in the minds of geneticists ever since the chromosome theory of heredity became established, and to which our answers are still partial and imperfect. How can cells containing identical nuclei, and consequently identical information at the level of genic DNA, nevertheless develop into such different types as a leaf cell, an epidermal cell, a large hair, a xylem cell, or a sieve tube?

Three general kinds of answers can be given to this question. The first would be that we are mistaken in our belief that all the cells of a higher plant or animal contain the same genetic information, but that, in fact, the genetic code is altered in a regular systematic fashion as cells and tissues develop. Although recent evidence from nuclear transplantation in frogs has suggested that in these animals irreversible changes in the nuclei may occur during development, much evidence from higher plants indicates that, in them, differentiated cells have retained all the genetic information that was present in the zygote at the time of fertilization, and are capable of regenerating whole plants. Entire plants have been obtained from single leaf cells of mosses,[7] epidermal cells of African violets *(Saintpaulia)*,[8] various kinds of cells in the carrot,[9] and sterilized cells of crown gall tissue in tobacco.[10] One phase of differentiation in all higher organisms tells us that a very high degree of cellular differentiation must be possible without changing the basic genetic code of information. This is the formation of eggs and sperm or pollen tubes. These are among the most highly specialized cells the organism produces, but at the same time they must contain unaltered genetic information which they pass on to the next generation.

The next type of answer is that of *differential gene action.* This would say that all the genetic information is present in every nucleus, but that each nucleus uses only that portion of the total information which enables it to develop the particular structure and function that is peculiarly adaptive to the tissue in which it resides. An analogy which might illustrate this answer would be to liken the nucleus to an enormous church organ containing thousands of pipes. If all the pipes were opened and played at once, the result would be merely a thunderous, cacophonous noise. But when the organ is serving its intended function, only a few pipes are played at a time, and the regular succession of combinations of pipes being played gives us a hymn, an oratorio, or a fugue.

This analogy can be carried farther, because we now have some idea of what in the living organism would correspond to the organ player and to the valves that open and close the pipes. Most of the evidence upon which these ideas are based comes from research on higher animals, but since it almost certainly is equally applicable to plants, I shall summarize it here. One of the most dramatic sequences of differentiation in higher organisms is that from larva to adult in the more highly evolved insects. Recent research on both moths and flies has shown that molting and the differentiation of the adult structures are triggered by the action of a steroid hormone known as ecdysone. This fact is of basic importance to geneticists, since flies possess in certain of their cells the well-known giant chromosomes in which the position of individual genes can be recognized. Furthermore certain of these gene loci become abnormally swollen or puffed at particular stages of development [illegible] and the pattern of puffing in different [illegible] of any particular tissue.[10] When ecdysone [illegible] is fed to cells in which chromosomes [illegible] a very high proportion of the label becomes [illegible] the puffed regions. This effect is counteracted [illegible] a known inhibitor of DNA-determined synthesis of RNA. Chemical analyses have shown that the RNA in puffs has all the characteristics of messenger RNA, which is produced by the DNA of the genes. This indicates that puffing is a sign of gene action in synthesizing messenger RNA, which, in turn, codes for protein. Equipped with this information Beerman and Clever[11] followed the changes in puffing at various sites on the salivary gland chromosomes of the midge *Chironomus* during the molting process. They found that each locus went through a characteristic series of changes. Some made puffs early, some late, some not at all. Clever and Carlson then injected ecdysone into *Chironomus* larvae and found that it caused puffing

Furthermore, each of the loci observed responded to a different concentration of the hormone. One locus responded to low concentrations of ecdysone, another locus to higher concentrations of the same hormone, and still others not at all. Other, unidentified factors in the larval cells were found also to control puffing.

This evidence suggests that the role of the "player" is played in part by hormones, some of which may control development by stimulating or inhibiting gene action. Varner and Chandra[12] have recently obtained evidence which indicates that the plant growth substance, gibberellin, promotes the germination of barley seeds by activating genes responsible for producing the enzyme alpha amylase. We cannot, however, jump to the conclusion that hormones are the only regulators in gene action in development. As Dr. Beermann[13] has pointed out, many kinds of differentiation are restricted to single initial cells, such as the bristles of *Drosophila* and the large epidermal hairs found on many plants. This suggests that generalized substances like hormones, which usually move through entire tissues or even the whole organism, can be responsible for controlling only some phases of gene action in development.

The evidence from puffing in the chromosomes of flies suggests that the ability of a gene to respond to a stimulus by synthesizing RNA and proteins depends upon the structure of the gene locus itself. The critical characteristics of gene structure need not, however, reside in the molecules of DNA. Recent studies of chromosomes in higher organisms[14] have shown that they contain, in addition to DNA, both RNA and proteins of various sorts. The attention of geneticists and biochemists has been attracted particularly to the basic proteins or histones. Studies by Huang and Bonner[15] on isolated pea nuclei, and particularly by Allfrey et al.[16] on extracts from calf thymus nuclei, have shown that DNA complexed with histone is unable to synthesize RNA in the

presence of RNA polymerase, but that removal of the histone activates the DNA. More recently, Allfrey et al.[17] have shown that, if the histone remains complexed with the DNA but acquires acetyl groups attached to the histone molecules, it no longer inhibits the activity of the DNA in serving as a template for RNA synthesis. Since acetylation of histones is reversible, the authors regard this process as a likely method for regulating gene action in the nucleus

These discoveries are far too recent for us to estimate their general significance at present. They need to be verified on a great variety of other kinds of chromosomes. Nevertheless, they do provide a possible biochemical basis for regulating gene action at the level of the gene itself Genetic evidence for such regulation is well known in bacteria[18] and maize[19,20] I believe it safe to predict that chemical "valves" for turning genes on and off will be clearly demonstrated in the very near future

Differentiation at the Level of Single Cells

These hormone and histone systems for regulating gene action, even if their existence becomes well established in a variety of organisms, will still be insufficient to explain many well-known features of cellular differentiation A conspicuous feature of development in higher plants, as well as in animal tissues, is the immediate and often striking differentiation of the two daughter nuclei from a single mitotic division, beginning as early as telophase A classic example is that of the pollen grain or microspore division in the developing anthers of flowering plants[21] At the telophase of this division the generative nucleus, which will form the two sperm nuclei, is already strikingly different from the vegetative nucleus, which will direct the metabolism of the pollen grain and pollen tube. This differentiation is associated with the position of the generative nucleus on the side of the pollen grain

which, at an earlier stage of development, was nearest to the center of the pollen tetrad. If, as a result of treatment with heat or cold, the pollen grain mitosis becomes reoriented so that the nuclei lie side by side and equidistant from the inner side of the grain, their differentiation into generative and vegetative nuclei fails to take place.[22] The most likely explanation of these facts is that, even before the microspore tetrads are formed, some kind of cytoplasmic gradient exists in the pollen mother cell between its outer and inner regions, and that some features of this gradient persist until the microspore division. Consequently, the daughter nuclei from this division, which must be genetically identical, react differentially to differences in their cytoplasmic environment. It is, of course, possible that this differentiation is produced by growth substances similar to the auxins, gibberellins, and kinins which are responsible for differential growth of plant organs, but we have absolutely no evidence that this is so. Even if it is, we are left with the most critical question of all: How can growth substances become unequally distributed within the confines of a single cell?

Differential mitoses similar to those found in pollen grains occur in a great variety of developing plant tissues. I have observed that in the leaf sheaths of grasses, particularly those high up on the stem, differential mitoses may lead not only to the differentiation of stomata but also, dependent upon the species, to one-celled hairs, two-celled hairs, or to pairs of characteristically differentiated cells, one with siliceous and the other with suberous walls (Fig. 40, A). The differentiation of cell initials which will later produce these different kinds of cells takes place in a region of the developing leaf sheath halfway between the most actively dividing intercalary meristem at the base of the sheath, and the zone of cell elongation that exists in its upper half. In the meristem itself, all mitoses are nondifferential, and the two daughter cells of a particular mitosis are indistinguishable until they divide again. In the



FIG. 40. A A portion of the outer (abaxial) surface of the devel-
[illegible]

Reproduced at a magnification of 500×

B An earlier stage of development of the same leaf sheath, showing formation of stomatal initials and of short cells which will give rise to siliceous–suberous pairs Same magnification

intermediate region, mitoses in every row of cells produce distally a cell which is smaller and has a denser cytoplasm, whereas the proximal cell from the same mitosis is larger, has a less dense cytoplasm, and may be vacuolate (Fig 40, B) Dependent upon the position of the cell row relative to the vascular bundles, the smaller distal cell may divide again to produce either a pair of guard cells or a siliceous–suberous pair, or may enlarge to produce a unicellular hair. The larger proximal cell always elongates and differentiates into a typical epidermal cell

These facts can be best explained by assuming that mitoses taking place in the meristem itself result in equal distribution of both nuclear and cytoplasmic elements, but that, in the post-meristematic divisions, certain elements of the cytoplasm become unequally distributed, so that nuclei in the distal end

of the cell receive a stimulus to further division or to differentiation of hairs, while the cytoplasm in the proximal end of the same cells is devoid of such stimulating elements. In our laboratory, we are studying the fine structure of these differentiating cells under the electron microscope in order to find out whether some of their visible organelles are unequally distributed during a differential division. Our first pictures show that plastids, mitochondria, golgi bodies, and membranes of the endoplasmic reticulum are present in about equal numbers in both of the daughter cells from a differential division. The conspicuous difference between these two cells is that in some instances the smaller one contains a higher concentration of particles which, after fixation with permanganate, look like minute granules. We suspect that these are ribosomes. This difference, however, was not found in all pairs of long and short cells. At present we can only say that recognizable differences in the fine structure of the cytoplasm are associated with the differential mitoses in the leaf sheaths, but that these do not involve the larger organelles.

The key facts that may provide a clue to understanding this phenomenon are first that differential mitoses occur only in the transitional region between actively dividing meristem and the zone of cell elongation, and second that the cytoplasm is always polarized so that the daughter cell that will later divide or differentiate most actively is in the distal position. One might suspect that a cytoplasmic gradient is being set up by some substance which is diffusing either upward or downward over the epidermis of the leaf, but as yet we have no evidence for this.

Genic Control of Cell Division Patterns

The fact that many differential mitoses give rise to one daughter cell which will divide again and one which will not,

tells us that patterns of cell division can be controlled in part by elements of the cytoplasm This fact will be of great importance in helping to explain the mode of action of such genes as those that control the formation and the shape of the flower and its parts. Since most of the significant steps in the evolution of higher plants have involved changes in the patterns of cell division, we cannot fully understand the processes of evolution until we have learned something about how these genes act.

The action of these genes differs sharply in one important respect from that of genes which are responsible only for the production of a particular kind of molecule, such as an amino acid or a pigment substance Genes responsible for the growth patterns must affect, in a very precise fashion, a succession of cell generations The genes for spurs on the petals of the columbine transmit information to a group of cells at the base of the petal which tells them to divide a definite number of times. The number of cells per spur is nearly the same not only in all five petals of a flower but also in the different flowers on a plant From the attachment of the petal to the nectary at the base of the spur, the number of cells in each row is between 300 and 320 in the red columbine of California This means that the gene complex responsible for spurs in this species tells the cells to go through eight or nine mitotic cycles, and to stop dividing when a definite number of cells has been produced Such action is best explained by assuming that these genes are of a regulator type and exert their effects specifically upon replication of DNA, the formation of mitotic spindles, or both

Because they are long-lived perennials and have very small cells, the species of columbine are not very favorable material for studies of gene action in development. Cultivated barley is a better subject for such investigations, since it is easily handled, it has relatively large cells and chromosomes, and its chromosomes are marked by a large number of recognized

gene loci, many of which affect the structure of the reproductive head or spike. We have been studying two of these gene loci and, in respect to one of them, have obtained some information about the way the gene alters the pattern of cell division.

The gene concerned is a dominant known as hooded, which has existed in collections of barley varieties for more than a hundred years. Its morphological effect is confined to the upper part of the fertile scale or lemma which subtends the floret and grain. It converts the awn or beard, which forms the upper part of this structure, into a hood-like organ that usually contains two extra rudimentary florets, connected by a small stalk or rachilla (Figs. 41 and 42). Although the hood is a complex structure consisting of many different parts, their development is controlled simultaneously by the action of a single gene. In F_2 progeny of crosses between hooded and awned varieties, a simple segregation of 3 hooded to 1 awned is regularly found.

The first visible action of the hooded gene occurs when the lemma primordium is about 300 micra long. At this stage, the cells on the adaxial or inner surface of the upper half of the primordium start to divide more rapidly than do corresponding cells of the awned genotype. These more rapidly dividing cells also change the orientation of their planes of cell division, so that they divide in three directions. The corresponding cells of the awned genotype all divide with their mitotic spindles parallel to the long axis of the primordium, thus giving rise to the long straight awn (Fig. 43, A and B). In hooded barley, cell division in three planes soon gives rise to an elevated dome or cushion, which in its histological structure resembles closely the axial meristem of the normal spike, from which the functional reproductive structures are differentiated (Fig. 43, C and D). The multiple end effects of the hooded gene can therefore be understood if we assume that



FIG 41 A portion of the mature head of hooded barley, showing the three spikelets at a node (4×)

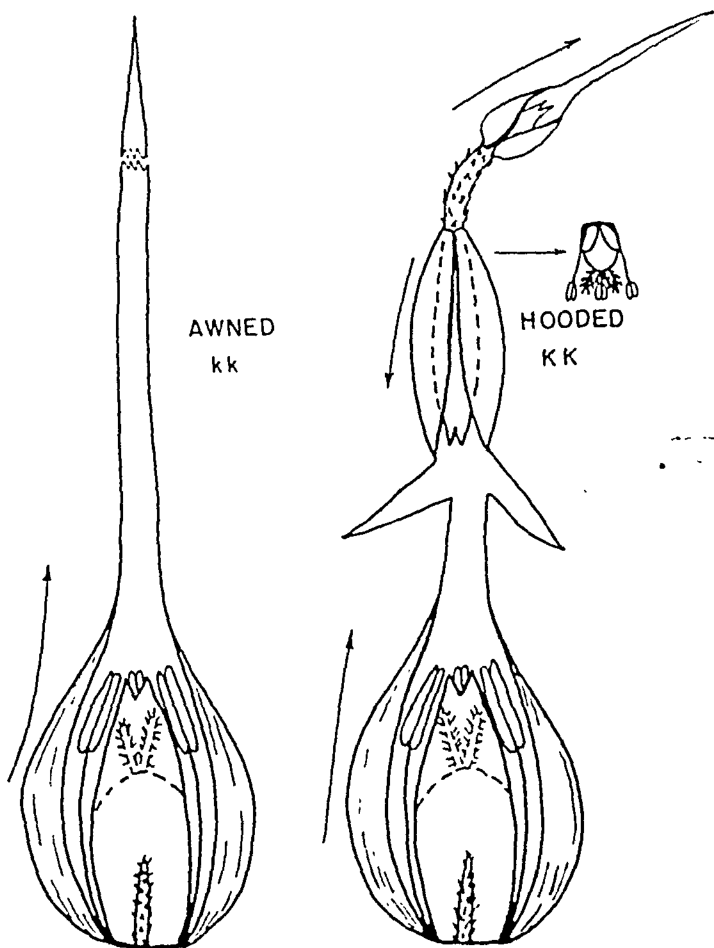


Fig. 42. The difference in structure between the lemma appendage of awned (kk) and hooded (KK) barley. Arrows show the position of developmental gradients.

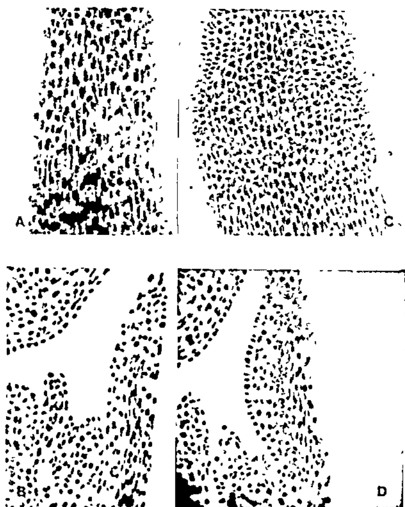


FIG 43 A and B Adaxial surface and cross section of the distal portion of a lemma primordium in awned barley
C and D The same in hooded barley All magnified 188×

it causes the cells in the upper half of the primordium to revert to an earlier state of ontogeny, so that cells descended from them go through a second cycle of organ differentiation. This shift in ontogenetic states is apparently the consequence of the accelerated mitotic cycle and the shift in orientation of mitoses from one plane to three.

By applying tritiated thymidine to spike primordia, Dr. Ezra Yagil, from our laboratory, has shown that the accelerated mitosis is associated with an increased rate of synthesis of DNA which is strictly localized in the region of the developing cushion. At a slightly later stage, the area where the cushion will develop stains heavily with pyronin and shows strong incorporation of label after application of tritiated cytidine (Fig. 44). These experiments suggest that the chemical action of the hooded gene, which shifts the ontogenetic state of the primordial meristem and brings about the new cycle of organ differentiation, is an increased rate of nucleic acid synthesis, first DNA then RNA.

Another experiment by Dr. Yagil provides further evidence to indicate that all the differences between the awn and the hood are brought about by a change in a single controlling process, which shifts the path of development into a new channel. If plants of the hooded genotype containing developing spikes are subjected to cold shocks of 5°C for ten days, the hood can be converted into an awn. This change is quantitative, and dependent upon the sensitivity of the plants, various degrees of intermediate conditions between hoods and awns can be obtained. The intermediate condition exists for all parts of the hood; as the overall shape becomes more awnlike, the various rudimentary parts of the extra florets become smaller or are suppressed entirely.

The experiments with cold shocks tell us also that the key process controlling hood formation takes place at the exact stage in development when the change in cell pattern can be seen in the lemma primordium. If cold is applied before the



FIG 44 Longitudinal sections of the apical regions of lemmas of awned and hooded barleys after treatment for 24 hours with tritiated thymidine at a strength of 1 microcurie per milliliter, showing nuclei in which incorporation of thymidine indicates synthesis of DNA during the period of exposure (233×)

primordia have reached the stage of cushion formation, they do not respond As might be expected, primordia exposed to cold after the cushion is well formed continue to make hoods regardless of the cold shock All these results are best explained by assuming that hood formation is controlled by a single key process which brings about an increased rate of nucleic acid synthesis, is temperature sensitive, and is suppressed by low temperatures

We must next ask ourselves, is this key process the primary action of the hooded gene? We have some evidence to indicate that it is not A few years ago we compared hooded and awned genotypes with respect to their pools of free amino

acids in seedlings six days old.[23] Although, at this stage, the two genotypes are completely indistinguishable morphologically and histologically, they show significant differences in the quantity of six amino acids, and these are much accentuated if the seedlings are grown in weak solutions of the antibiotic chloramphenicol. At other stages of development, significant differences have been found in carotenoid content, which at some stages is higher in the awned genotype and at others in the hooded genotype. These differences in chemical composition at stages when no morphological differences are evident suggest that the primary action of the hooded gene may be taking place at many, and perhaps all stages of development, but that it has a significant effect on cell behavior only in the lemma primordium. In another mutant, known as *agropyroides,* which has morphological effects confined to the upper part of the reproductive spike, differences in free amino acid content can also be found at stages when no morphological differences are evident.[24]

These examples suggest that a third answer can be given to the question of how cells with identical gene contents can produce very different morphological structures. Some genes may be acting at subthreshold levels throughout development, but have a significant effect on cell behavior only when their action is combined with that of other genes to produce a particular type of cellular environment. In this connection, the fact is perhaps significant that the stage of cushion formation in the hooded genotype is one in which the lemma primordium in any type of barley is growing at its maximum rate, and presumably a maximum supply of raw materials is entering the primordium. We might postulate that the hooded gene is able to increase the rate of nucleic acid synthesis significantly only when a maximum supply of raw material for these syntheses is present.

All the available facts about the action of the hooded gene make sense on the basis of the following hypothesis. One of

the secondary effects of this gene is to increase the rate of nucleic acid synthesis, both RNA and DNA, in meristematic tissue which is growing at its maximum rate. In the lemma primordium, this brings about simultaneously an increased rate of cell proliferation at the expense of cell enlargement and a change in orientation of cell divisions from one plane to three. In this way, a new reproductive meristem is formed, which has all the properties of such a meristem, including the ability to differentiate new floral parts A further postulate is that the increased cell proliferation is temporary, and that a second cycle of organ differentiation takes place as the rate of cell division starts to decline.

If this hypothesis is correct, it may lead to a new understanding of how differences in genes can give rise to differences in form. Particularly significant in this connection is the suggested relationship between rate and direction of cell division As exemplified by the differences between awned and hooded, this relationship can be expressed as follows (Fig. 45) If the rate of mitosis is lagging behind that of cell

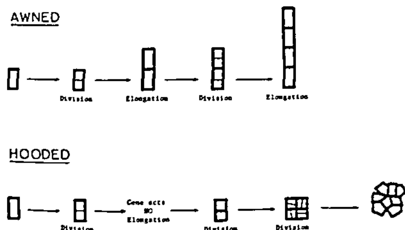


FIG. 45 The relationship between cell elongation and orientation of cell division in awned and hooded barley

enlargement, so that cells are becoming larger at each successive mitotic prophase, then the orientation of cell division will be governed by forces which govern the growth of the tissue or organ as a whole and control the direction of cell enlargement. If, on the other hand, the tempo of successive cell divisions equals or exceeds that of cell enlargement, so that cells are staying the same size or becoming smaller at each successive prophase, then the direction of cell division will be regulated by the tempo of the mitotic rhythm, and, with a maximum rate of this rhythm (i.e. a minimum duration of the interphase between mitoses), the orientation of cell division will be in three planes.

If this relationship exists generally in plant tissues, it will go a long way toward explaining how the action of genes in changing the rates of processes of cellular metabolism can affect the shape of organs at the macroscopic level. It would tell us that any change in metabolism which would tend to reduce the length of the interphase between mitoses below a critical level would automatically affect the orientation of cell division, and consequently the shape of the resulting structure. Conversely, in tissues which normally are dividing in two or three planes, an increase in the length of the interphase between mitoses should tend to reorient mitotic spindles in a single plane. This point is so important that evidence bearing upon it should be obtained from other systems. We have obtained such evidence from the developing stomata in the leaves of barley.

In a grass leaf, the stomata are arranged in linear rows, which extend the length of the leaf, and contain hundreds of stomata. Each stomatal opening is flanked by two guard cells and two subsidiary cells. The way stomata develop in barley has been described elsewhere.[25] Here I should like only to draw attention to the unique plane of orientation of the mitotic division which produces the two guard cells. In all the previous mitoses which have taken place in the stomatal row,

and in all divisions of epidermal cells between stomatal rows as well as those in the leaf parenchyma, the spindle is oriented parallel to the long axis of the leaf. The final mitosis of the guard cell–mother cell to form the two guard cells stands out in having its spindle oriented transversely to the leaf axis. How does this happen?

The orientation of other mitoses parallel to the axis of elongation of the leaf can be explained by a well-established tendency of meristematic cells to divide in the direction of the greatest rate of cell elongation,[26] which, in grass leaves, is always parallel to the long axis of the leaf. As a result, we might explain the exceptional orientation of the stomatal guard cell division on the basis of the observed fact that the guard cell–mother cell elongates less before it divides than do the surrounding cells, and consequently is elongating little or not at all in a direction parallel to the leaf axis. If, therefore, we could cause the guard cell–mother cell to elongate parallel to the long axis of the leaf before dividing, we should be able to change the plane of orientation of its spindle and make it parallel to the long axis of the leaf.

We have done this in two ways. In collaboration with Dr. S. S. Shah,[27] barley seedlings were immersed for one hour in a solution of 0.2 molar 2-mercaptoethanol. As demonstrated by Mazia and Zimmerman,[28] this compound interferes with the formation of the mitotic spindle but has no effect on DNA replication. In our seedlings, it blocks mitoses temporarily, but, after the dosage we used, growth is resumed after about 16 hours, and the seedlings develop normally. In leaf meristems fixed 18-24 hours after treatment, small numbers of mitoses of guard cell–mother cells are now oriented parallel to the long axis of the leaf. These can be increased significantly by growing the seedlings in low concentrations of gibberellin, which increases the amount of cell elongation.

A much more striking result has been obtained from

sheaths of stem or culm leaves, by the simple process of removing them from the plant and placing them for 24 hours on wet filter paper in a petri dish. Under these conditions, mitosis continues for a short while, but, 8 hours after removal, mitotic divisions have almost ceased. At about 16 hours after removal they are more frequent, while at 20 and 24 hours after removal the numbers of guard-mother cell mitoses seen in a fixed leaf are, respectively, 47 and 32 per cent of what they were at the time of removal. Furthermore, about 88 per cent of the mitoses that take place after the quiescent period have spindles oriented parallel to the long axis of the leaf, and form guard cells not side by side but proximal and distal to each other (Fig. 46).



FIG. 46. A portion of the outer (abaxial) epidermis of the developing sheath of the uppermost culm leaf of barley, after removal and placement for 24 hours on wet filter paper in a covered petri dish. At right, a normal developing stomatal complex. At left, several complexes in which reorientation of the spindle at right angles to its normal plane has produced a proximal-distal arrangement of the guard cells (920×).

This effect has been obtained only in the sheaths of culm leaves. When blades of these leaves or sheaths of basal leaves are treated in the same way, few or no proximal-distal mitoses are obtained. This is associated with the fact that, in these organs, few of the guard cell-mother cells resume mitosis after the quiescent period.

These experiments provide further evidence in favor of the hypothesis that, in plants, shapes of organs are determined to a large extent by relative rates of mitosis and cell enlargement occurring in meristematic shoot apices and very young primordia. The genes that mold the different forms of the columbine flower, as well as the delicate tracery of a fern leaf, the regal symmetry of a lily flower, and the fantastic sculpture of an orchid, may well do their work by commanding an army of cell nuclei to execute complex sequences of divisions and replications at ever-changing but carefully and precisely controlled rates. The exploration of such a possibility has great fascination for a plant evolutionist and should provide problems for botanists to investigate for many years to come.

Our research has been aided over the years by grants from the National Science Foundation. The research described in the later part of this paper was conducted with the aid of National Science Foundation grant HeB 2276.

REFERENCES

1. D. E. Koshland, Jr., Correlation of Structure and Function in Enzyme Action. *Science 142*, 1533 (1963).
2. R. F. Mallina, T. R. Miller, P. Cooper, and S. C. Christie, Surgical Stapling. *Sci. American 207* (4), 48 (1962).
3. J. Monod, J.-P. Changeux, and F. Jacob, Allosteric Proteins and Cellular Control Systems. *J. Mol. Biol. 6*, 306 (1963).
4. J. A. Weir, Sparing Genes for Further Evolution. *Proc. Iowa Acad. Sci.*, 53, 313 (1946).
5. C. L. Markert, Epigenetic Control of Specific Protein Synthesis in Differentiating Cells. in *Cytodifferentiation and Macromolecular Synthesis*. M. Locke, Ed. 21st Growth Symposium. Academic Press. New York (1963). p. 65.
6. W. Prazmo, Genetic Studies on the Genus *Aquilegia L.* I. Crosses between *Aquilegia vulgaris L.* and *Aquilegia ecalcarata* Maxim., *Acta Soc. Bot. Polon. 29*, 57 (1960).

7. E. Zepf, Über die Differenzierung des Sphagnumblattes, *Z. Bot.*, *40*, 87 (1952).

8. A. H. Sparrow, R. C. Sparrow, and L. A. Schairer, The Use of X-rays to Induce the Somatic Mutations in *Saintpaulia*, *African Violet Mag.*, *13*, 31 (1960).

9. F. C. Steward, Growth and Development of Cultured Plant Cells, *Science*, *143*, 20 (1964).

10. A. C. Braun, A Demonstration of the Recovery of the Crown-Gall Tumor Cell with the Use of Complex Tumors of Single-Cell Origin, *Proc. Nat. Acad. Sci. U.S.*, *45*, 932 (1959).

11. W. Beermann and U. Clever, Chromosome Puffs, *Sci. American*, *210* (4), 50 (1964).

12. J. E. Varner and G. R. Chandra, Hormonal Control of Enzyme Synthesis in Barley Endosperm, *Proc. Nat. Acad. Sci. U. S.*, *52*, 100 (1964).

13. W. Beermann, Cytological Aspects of Information Transfer in Cellular Differentiation, *Am. Zoologist*, *3*, 23 (1963).

14. H. Ris, Ultrastructure and Molecular Organization of Genetic Systems, *Can. J. Genet. Cytol.*, *3*, 95 (1961).

15. R. C. Huang and J. Bonner, Histone, a Suppressor of Chromosomal RNA Synthesis, *Proc. Nat. Acad. Sci. U. S.*, *48*, 1216 (1962).

16. V. G. Allfrey, V. C. Littau, and A. E. Mirsky, On the Role of Histones in Regulating Ribonucleic Acid Synthesis in the Cell Nucleus, *Proc. Nat. Acad. Sci. U. S.*, *49*, 414 (1963).

17. V. G. Allfrey, R. Faulkner, and A. E. Mirsky, Acetylation and Methylation of Histones and Their Possible Role in the Regulation of RNA Synthesis, *Proc. Nat. Acad. Sci. U. S.*, *51*, 786 (1964).

18. F. Jacob and J. Monod, Genetic Regulatory Mechanisms in the Synthesis of Proteins, *J. Mol. Biol.*, *3*, 318 (1961).

19. B. McClintook, Some Parallels Between Gene Control Systems in Maize and in Bacteria, *Am. Naturalist*, *95*, 265 (1961).

20. R. A. Brink, Paramutation at the R. Locus in Maize, *Cold Spr Harbor Symp. Quant. Biol.*, *23*, 379 (1958).

21. K. Sax and K. W. Edmonds, Development of the Male Gametophyte in *Tradescantia*, *Bot. Gaz.*, *95*, 156 (1933).

22. K. Sax, The Effect of Temperature on Nuclear Differentiation in Microspore Development, *J. Arnold Arboretum*, *16*, 301 (1935).

23. IGOR SARKISSIAN, S. S. SHAH, and G. L. STEBBINS, Differences in Free Amino Acid Content of Seedlings of Awned and Hooded Barley, and Their Alteration by Chloramphenicol Treatment, *Proc Nat Acad Sci U S*, *48*, 1513 (1962)

24. R. T. WIJEWANTHA and G. L. STEBBINS, Developmental and Biochemical Effects of the *Agropyroides* Mutation in Barley, *Genetics*, *50*, 65 (1964)

25. G. L. STEBBINS and S. S. SHAH, Stomatal Development in the Leaf Epidermis of certain grasses, *Develop Biol*, *2*, 477 (1960)

26. E. W. SINNOTT, *Plant Morphogenesis*, McGraw-Hill, New York (1960)

27. G. L. STEBBINS, S. S. SHAH, D. JAMIN, and P. JURA, Reorientation of the Mitotic Spindle of Stomatal Guard Cell Divisions in *Hordeum vulgare*, *Am J Bot*, *54*, 71 (1967)

28. D. MAZIA and A. M. ZIMMERMAN, SH compounds in Mitosis. II. The Effect of Mercaptoethanol on the Structure of the Mitotic Apparatus in Sea Urchin Eggs, *Exp Cell Res*, *15*, 138 (1958)

10. LIQUEFIED NATURAL GAS
A NEW SOURCE OF ENERGY

By C. M. Sliepcevich
University of Oklahoma

I. SHIP TRANSPORTATION

The population explosion accompanied by an ever-rising standard of living is dramatically reflected in the world's gigantic demand for more energy As aptly stated by Schurr [1]

> Modern man has made himself largely by burning fuel The supply of fuel appears to be almost inexhaustible, and a high level of fuel consumption is not a prerequisite of development but a result of it

Sporn[2] emphasizes that the availability of energy does not guarantee industrial development, rather, the capacity to consume energy, not to produce it, is the factor Sporn's point is directed toward misconception by world leaders following the 1955 Geneva Conference on the Peaceful Use of the Atom They believed that the advent of the atom as an energy source would not only provide a cheap, inexhaustible source of energy but it would also convert overnight even the least developed countries into economies and civilizations comparable to those of the United States and Western Europe What they failed to recognize was that the mere presence of available energy could accomplish little without the capital equipment to utilize it

Man's social, economic, and technical progress can be measured by his progress in using energy effectively.[3] The United States serves as a vivid example. With only 6 per cent of the world's population, it consumes about 35 per cent of the world's commercial energy (excluding wood, animal power, etc.); whereas India with about 15 per cent of the world's population uses only 1.5 per cent. A study of our annual energy consumption reveals the economic and technical history of the United States as it emerged from an agricultural economy during the colonial period to the present age of spectacular industrial growth.[4]

As late as 1880, wood was our major source of energy, but by 1890 coal was supplying as much energy as wood. With the growth of electricity, coal quickly replaced wood and by 1910 was accounting for 75 per cent of the energy. Coal consumption peaked in 1918 but, with the advent of the automobile, liquid petroleum products began to rise rapidly in prominence. After World War II, as a result of the development of welded, seamless pipe which made long-distance gas transmission lines feasible, combined with an enormous upsurge in new residential housing, natural gas further depressed the position of coal in the world's energy picture. The growth in energy consumption and the shift in distribution of energy sources over the past three decades are summarized in Table 7.

It will be noted from Table 7 that natural gas experienced the greatest rate of growth, 350 per cent; most of it has occurred in the United States. While the growth of liquid fuels in the last few years has stagnated, natural gas has maintained about a 7 per cent annual growth rate that could cause it to surpass liquid fuels as the major source of energy.

The world production and proved reserves of natural gas for 1961 are presented in Table 8.

The amount produced as shown in this table does not include the vast quantities of gas that are wasted by flaring in-

TABLE 7. CHANGES IN ENERGY CONSUMPTION BETWEEN 1929 AND 1960*

	World		United States	
	1929	*1960*	*1929*	*1960*
Total consumption in million metric tons of hard coal equivalents†	1,711	4,235	777	1,448
Distribution in per cent				
Solid fuel	79 8	52 3	67.9	24.6
Liquid fuel	14.9	31 1	22 3	40 8
Natural gas	4 5	14 6	9.3	33 3
Hydroelectric	0 8	2 0	0 5	1.3

**Sci Am*, *209*, 114 (Sept 1963)
†The thermal value of all energy sources is converted to equivalents of hard coal

TABLE 8 WORLD PRODUCTION AND RESERVES OF NATURAL GAS*
(Trillions of Cubic Feet)

	Cum Prod Through 1961	*Prod 1961*	*Proved Reserves End 1961*
North America	242 8	14 6	321 0
Middle East	10 0	1 14	178 3
Iron Curtain	19 4	2 65	83 4
Africa	0 3	0 08	54 5
South America	19 5	1 7	44 8
Far East	2 8	0 17	19 6
Europe	3 0	0 55	19 1
Total Free World	278 4	18 24	637 3
Total Iron Curtain	19 4	2 65	83 4
Total World	297 8	20 9	720 7

*Summarized from *Oil and Gas J*, *60*, 75 (March 12, 1962)

to the atmosphere, particularly in the Middle East and South America. The proved reserves, particularly in Europe, are subject to substantial revision as of 1965, because of the large amount of gas discovered in Holland in 1959 but only

recently disclosed as the third largest gas field in the world. About 70 per cent of the gas produced in 1961 was in the United States. The distribution among users and the corresponding revenues are summarized in Table 9.

It is interesting to observe in Table 9 that the residential customer provides more than 55 per cent of the annual revenue, which approaches 6 billion dollars. The revenue produced by the residential customer amounts to 10 cents per unit of thermal energy (equivalent to 100 cu ft of gas), whereas the industrial customer provides only 3.5 cents per therm. As will be explained later, it is this superficial bargain price that industry receives—based on interruptible supply—that actually catalyzed the development of the liquefaction and transportation of natural gas on an international scale.

To judge from its effective utilization in the United States, particularly in the last twenty years, natural gas plays a prominent role in an expanding, industrial economy. The situation in the United States has been unique in that it not only has about half of the world's proved reserves of natural gas but also that the gas can be distributed competitively from the major producing areas in the southwest[1] to all the consuming centers by means of pipelines. On the other hand, some of the highly industrialized nations like Japan are (as yet) neither blessed with indigenous gas reserves nor can they be reached by pipeline from major producing areas. Since it is impractical to transport natural gas as a gas in bulk form, one obvious solution would be to liquefy the natural gas and transport it as such in ocean-going tankers. The advantage in liquefaction is that roughly 600 cu ft of natural gas at atmospheric pressure shrinks to 1 cu ft of liquid. Other techniques which have been considered are transporting in

1. Natural gas has been found under less than 1 per cent of the land area of the United States. This gas is transmitted in pipelines (10 to 36 inches in diameter) to consuming markets as far as 2,000 miles away at pressures between 200 and 1000 psi.

TABLE 9 U.S SALES OF NATURAL GAS DURING 1961*

Category	*Average Number of Customers*	*Thousands of Therms*†	*Revenues*
Residential	29,105,000	31,790,900	$3,183,981,000
Commercial	2,392,000	9,599,200	760,033,000
Industrial	136,000	46,844,400	1,622,595,000
Other	38,000	4,815,100	165,236,000
Total	31,671,000	93,079,600	$5,731,815,000‡

*[illegible]

gain in therms

the gas phase at high pressure, by absorption in liquids, by adsorption on solids, and by reversible chemical combination.[5] A comparison of these various systems is given in Table 10.

Since it is impractical to store or transport large volumes of gas at any pressure above atmospheric, the most competitive alternative to liquefaction is adsorption on fuller's earth Other solids such as carbon can be used, but its adsorptive capacity is less than 10 per cent of fuller's earth Considerable

TABLE 10 COMPARISON OF VARIOUS SYSTEMS FOR STORING AND TRANSPORTING NATURAL GAS*

System	*Cu Ft of Volume Required to Contain 1000 Cu Ft of Gas*	*Conditions*	
		Temperature °F	*Pressure, psi*
Liquefaction	1 6	−250	14 7
Adsorption (fuller's earth)	5 0	−250	14 7
Hydrate	5 9	35	150
Absorption (in propane)	6 5	− 52	600
Compression	25 0	100	600

*Chem Eng Progr, 58, 47 (Nov 1962)

developmental work was done, including the operation of a pilot plant in Warren, Pa., by the Floridin Co. and J. F. Pritchard Co. in 1949–50, on fuller's earth adsorption.[6] However, because of its 3:1 disadvantage on containment volume as compared to liquefaction, and the fact that the same temperatures and pressures are required, this process has never been commercialized.

At this point, it may be well to review some of the more pertinent properties of natural gas. It is found in porous, subsurface, imperviously capped formations in various parts of the world (see Table 8). Some wells are more than three miles deep. Natural gas is frequently found with petroleum; about one third of the U.S. production comes from oil wells. Natural gas in the United States generally contains between 80 and 95 per cent methane;[2] the balance includes ethane, propane, butane, pentane, etc. Small amounts of carbon dioxide, nitrogen, helium, water vapor, and hydrogen sulfide are present in most natural gases. Natural gas, along with petroleum, oil shale, natural gas liquids, coal, and lignite, is classed as a fossil fuel.

Where the natural gas is composed primarily of methane, much of the characteristic behavior of natural gas can be predicted from the properties of methane, which are summarized in Table 11.

Since the critical temperature of methane is −116.5°F, it cannot be liquefied at any pressure, however great, above this temperature. The liquefaction temperature or boiling point at one atmosphere pressure is −258.68°F. This boiling point is compared with other gases, which are commonly liquefied, in Table 12.

Liquid methane is a colorless, clear liquid that resembles liquid air; its density is about one half that of liquid air. Because it possesses superior wetting characteristics, liquid

2. In the Middle East, much of the gas contains less than 50 per cent methane.

TABLE 11. PROPERTIES OF METHANE

Formula	CH_4
Molecular weight	16.042
Gas density at 60°F and 1 atm in lb/cu ft	0.0424 (0.679 gm/liter)
Specific gravity at 60°F and 1 atm (air = 1)	0.555
Critical temp., °F	−116.5°F (−82.5°C)
Critical press., psi abs	673.1 (45.8 atm.)
Boiling point at 1 atm press.	−258.68°F (−161.5°C)
Freezing point at 1 atm press	−296.46°F (−184°C)
Density of liquid at −263°F in lb/cu ft	25.9 (415 gm/liter)
Heat of vaporization at boiling point in Btu/lb	219.7 (122.1 cal/gm)
Net heat of combustion at atm press in Btu/cu ft	911 (21,240 Btu/lb or 11,800 cal/gm)
Flammability limits volume per cent in air	
Lower	5.0
Higher	15.0

TABLE 12. BOILING POINT OF SOME COMMON GASES

Substance	*Normal Boiling Point*
Ammonia	− 28.1
Freon 22	− 41.4
Propane	− 43.7
Carbon dioxide	−109.3 (sublimes)
Ethane	−127.6
Ethylene	−154.8*
Freon 14	−198.4
Methane	−258.7
Oxygen	−297.3
Nitrogen	−320.5
Hydrogen	−423.0
Helium	−452.1

*Temperatures below −150°F are usually considered cryogenic in the trade

methane produces a more severe irritation in contact with the human skin than liquid air does. It is this characteristic wetting property that may serve a useful purpose in cryogenic surgery as a replacement for liquid nitrogen.

Historical

The commercial liquefaction of natural gas dates to a small plant that was built in West Virginia in 1910 to compress natural gas, refrigerate and separate what was called liquid natural gas, mostly ethane and propane, which was bottled and sold locally. A patent application had been made by Cabot as early as 1914 for the liquefaction, storage, and barge transportation of liquid natural gas,[7] and in 1917 a patent was issued (U.S. Pat. 1,225,574) to him covering the apparatus for condensing natural gas under high pressure and cooling.

In 1917, during World War I, the United States Government commissioned the Linde Co., working in cooperation with the Bureau of Mines, to construct a plant in Fort Worth, Texas, for extracting helium from natural gas. The helium was to be used in dirigibles for the Allies. By the time enough helium (90 per cent purity) had been produced to fill one dirigible and was readied for shipment overseas from New York, the Armistice was signed.[8, 9] The process of recovering helium was based on liquefaction of natural gas. Some ten years earlier, Professors H. P. Cady and D. McFarland at the University of Kansas discovered that many natural gases contained around 1 per cent helium.

After the war, the government, under the jurisdiction of the navy, constructed a larger extraction plant at Fort Worth in 1921. In 1925, Congress placed all helium activity under the U.S. Bureau of Mines at Fort Worth. When the gas field near Fort Worth played out, the plant was closed and dismantled. A new one was erected near Amarillo, which went into operation in 1929 and has continued to produce helium ever since.

In the early 1920s, patents were issued on insulated containers for river barges, suitable for transporting liquefied gas.[10] In 1937, two patents were issued (U.S. Pat. 2,082,189 and 2,090,163) to L. Twomey on methods of liquefaction, storage, and delivery of liquefied natural gas through distributing lines.[11]

In 1937, H. C. Cooper,[3] then president of the Hope Natural Gas Co., became interested in liquefaction of natural gas with the result that a pilot plant was erected at Cornwell compressor station of the Hope Natural Gas Co of West Virginia in 1940. The liquefaction capacity was 300,000 cu ft of natural gas per day into a cork-insulated storage container for 1 million cu ft of gas (equivalent to 14,500 gal of liquid) Because of the successful operation of this plant, it was used as the basis for the design of a larger installation at Cleveland [10]

The Cleveland Natural Gas Liquefaction Plant of the East Ohio Gas Company went into operation on January 29, 1941 Total construction costs were $1 25 million This plant was known as a *peak-shaving plant* since its purpose was to liquefy surplus natural gas from the pipeline during the periods of low customer demand in the summer and to regasify the liquid from storage to supplement the pipeline gas during peak demands in the winter This plant was the first and only one of its kind in the United States, however, at least two liquefied natural gas, peak-shaving plants were scheduled to go on stream in 1965

The Cleveland plant had a capacity to liquefy 4 million cu ft of gas a day, to store a total of 150 million cu ft of gas as a liquid (equivalent to 1 8 million gal liquid) in three cork-insulated (3 ft thick) spherical tanks (57 ft in diameter), and to regasify the liquid at a daily rate of 72 million cubic feet After three years of successful operation, a fourth tank was

3 About this same time Egerton in England was trying to promote the separation and storage of liquid methane by the British Gas Industry to meet seasonal variations in demand on manufactured gas [12]

installed in 1944. This tank was a vertical cylinder, 70 ft in diameter and 43 ft high, surrounded by 3 ft of rock-wool insulation. Its capacity was 90 million cu ft of gas (equivalent to 1.1 million gal of liquid). Eight months after installation, the new cylindrical tank failed on October 20, 1944. Because of inadequate dikes, liquefied natural gas (hereafter referred to as LNG) flowed over the ground surface and into the sewers of the city. The resulting explosion and fire caused widespread destruction ($6.8 million) and loss of lives (128). A team of investigators from the Bureau of Mines observed that, among several possibilities, the most likely cause of failure was the improper selection of metal. The 3.5 per cent nickel steel which was used is not considered adequate for this service even by present-day standards of greatly improved fabrication techniques. The damage after tank rupture was attributed to the inadequacy of the earthen dikes. Despite the magnitude of the disaster, a significant conclusion by the investigation team is quoted from the Bureau of Mines report:[13]

> Regardless of the cause of the disaster at the liquefaction, storage, and regasification plant of the East Ohio Gas Company, the application of the system for liquefying and storing large quantities of natural gas is not invalidated, provided proper precautions are taken.

Nevertheless, the Cleveland plant was not operated again.

In 1947, Dresser Industries Ltd. of Dallas, Texas, designed and constructed a plant near Moscow, Russia, at a cost of 6 million dollars. This plant had a liquefaction rate of 4.5 million cu ft of gas per day (comparable to the Cleveland plant) with an equivalent gaseous storage volume of 162 million cu ft. Little information is available on the operating record of this plant other than it has been giving continuous, satisfactory service. Although the original purpose of the plant

another barge for transporting to Chicago. Another feature of the plan was to utilize the refrigeration contained in the LNG, during regasification at Chicago, to freeze and preserve the various products from the stockyards operation.

After considerable preliminary investigation, research, and development, construction of the barge-mounted liquefaction plant and the transport barge was undertaken in 1954 at Ingall's Ship Yard in Pascagoula, Miss. About this time, Prince decided it would be desirable to have a close working association with a company experienced in gas processing. After talking to several prospects, he found an understanding ear in E. F. Battson, senior vice-president of Continental Oil Co. Continental Oil took a one-year option on a possible joint venture with the Stock Yards group. A task force composed of Continental Oil Co. personnel and several consultants was organized under the direction of J. A. Murphy, who at that time was serving as Battson's technical advisor. This group not only made a detailed evaluation of the work done by the Chicago Stock Yards Research Division but they also carried out an independent, economic and engineering feasibility study. These studies concluded that the Mississippi barge venture was not economical, but that ocean transportation of LNG from gas-surplus countries to gas-deficient countries offered a very attractive potential. With this understanding, Continental Oil exercised its option to a joint venture, and in 1955 Constock International Methane Ltd. was organized with equal ownership by Continental Oil and Union Stock Yards and Transit companies. The name Con-stock was obviously derived from the parent companies.

Even though the original barging scheme was abandoned, it was decided to utilize both barges as pilot plants for demonstrating the technical feasibility and obtaining valuable design data for a commercial venture. The barges were completed in late 1955 and were then transferred to Bayou Long, La., for extensive testing throughout 1956.

generating power for the Chicago Stock Yards, and its dynamic chief executive William Wood Prince, to break the ice and take the bold plunge. Armed only with his self-imposed motto of awareness: "Remember Cleveland," Prince rolled up his sleeves and went to work on LNG in 1951.

The Constock Breakthrough

Most of the gas supplied to industry falls into the so-called interruptible category. Contracts with gas companies recognize that domestic consumers have first priority on the supply. In periods of severe weather, the supply to industry frequently has to be restricted. For this reason, industry must maintain standby facilities such as coal, liquid fuels, or manufactured gas to carry them through the period of interrupted service. However, industry is willing to accept this inconvenience so long as the gas companies will grant them a bargain price for the gas when there is a surplus during the summer months. As was noted in the comments related to Table 9, the average price paid by the industrial customer is about one third as much as that of the domestic customer per unit of thermal energy.

According to one source,[17] "a Chicago gas company injudiciously tried to raise the price of the natural gas (interruptible) it was selling to a power company controlled by William Wood Prince," then president of Union Stock Yard and Transit Co. and a managing trustee of the thirty-company Prince trust. It was then—in 1951—that Prince and one of his consultants, Willard Morrison, conceived the idea of liquefying natural gas on the Gulf Coast and barging it up the Mississippi River to Chicago. The original plan, which was handed to his Chicago Stock Yards Research Division to develop, was to construct a barge-mounted liquefaction unit which could move around to nearly depleted or remote gas wells (where the cost of gas is very low) and to liquefy the gas directly into

another barge for transporting to Chicago. Another feature of the plan was to utilize the refrigeration contained in the LNG, during regasification at Chicago, to freeze and preserve the various products from the stockyards operation

After considerable preliminary investigation, research, and development, construction of the barge-mounted liquefaction plant and the transport barge was undertaken in 1954 at Ingall's Ship Yard in Pascagoula, Miss About this time, Prince decided it would be desirable to have a close working association with a company experienced in gas processing After talking to several prospects, he found an understanding ear in E. F Battson, senior vice-president of Continental Oil Co. Continental Oil took a one-year option on a possible joint venture with the Stock Yards group A task force composed of Continental Oil Co personnel and several consultants was organized under the direction of J A Murphy, who at that time was serving as Battson's technical advisor This group not only made a detailed evaluation of the work done by the Chicago Stock Yards Research Division but they also carried out an independent, economic and engineering feasibility study. These studies concluded that the Mississippi barge venture was not economical, but that ocean transportation of LNG from gas-surplus countries to gas-deficient countries offered a very attractive potential With this understanding, Continental Oil exercised its option to a joint venture, and in 1955 Constock International Methane Ltd was organized with equal ownership by Continental Oil and Union Stock Yards and Transit companies The name Con-stock was obviously derived from the parent companies

Even though the original barging scheme was abandoned, it was decided to utilize both barges as pilot plants for demonstrating the technical feasibility and obtaining valuable design data for a commercial venture The barges were completed in late 1955 and were then transferred to Bayou Long, La., for extensive testing throughout 1956

While these tests were underway in Bayou Long, Constock concentrated on a crash program of research, development, and engineering analyses on all phases of the project, directed toward the commercial venture. Included were innovations in gas processing and liquefaction techniques, material evaluation and development, ship designs, cargo handling, storage tanks, etc. The most amazing aspect of the program was the way it was accomplished by its subsidiary, Constock Liquid Methane Corp. They operated with a skeleton staff directed by J. A. Murphy and housed in less than 500 square feet of space in a New York office building. Specific research and development assignments were doled out to the laboratories of the parent companies in Chicago and Ponca City, Okla. In addition, several consultants from universities were employed on a part-time basis, primarily to translate the research results into design criteria for practical applications. The bulk of the engineering design and construction was handled by J. F. Pritchard Co., Kansas City, Mo. (gas processing, liquefaction, and plant construction); Gamble Brothers, Inc., Louisville, Ky. (wood and insulation specialists); J. J. Henry Co., N.Y. (naval architects and marine engineers); and A. D. Little, Inc., Cambridge, Mass. (storage and cargo-handling methods).

By the spring of 1947, complete designs, specifications, and drawings for the liquefaction plant, tanker, and terminal facilities had been completed. Comprehensive analyses of potential gas sources and markets were also made. It was therefore possible to establish the economics for LNG shipments between a variety of ports all over the world. At this stage, a number of foreign countries, among which were England, Germany, France, Italy, Sweden, and Japan, expressed interest in importing LNG. By the fall of 1957, the British Gas Council made a declaration of intent to import LNG at a rate equivalent to 100 million cu ft of gas per day, which amounted to about 10 per cent of their total gas con-

sumption. However, [illegible] until after [illegible] made between [illegible] Even though the Constock ship [illegible] appeared to be sound, there was as yet no proof that LNG could be transported overseas by tanker, particularly in rough weather.

Constock agreed to erect liquefaction and land-storage facilities on the Calcasieu River near Lake Charles, La., while the British Gas Council supplied the unloading and terminal storage facilities at Canvey Island near London. Constock and the Gas Council agreed to share in the costs of converting a dry-cargo tanker; for this purpose, a new company, British Methane, Ltd., was formed to own and operate the ship.

Lake Charles facilities. The liquefaction plant, land storage, and loading terminal were erected on a twenty-acre site and placed in operation during the later part of 1958. The barge-mounted liquefaction unit, which was used in the Bayou Long tests, was moved to Lake Charles. A simplified flow sheet of the modified Claude (expander) cycle used on the barge is shown in Figure 47. Although this cycle is relatively inefficient, it has the advantage of being lighter and more compact, and therefore more adaptable to installation on a barge where space is limited. It is interesting to note that this pilot liquefaction unit with a rated liquefaction capacity of 7 million cu ft of gas per day was 17 times larger in capacity than the Cleveland or Moscow plants of the 1940s.

The liquid from the plant was stored in a tank having a capacity of 1.47 million gal (equivalent to 120 million cu ft of gas). Up until 1961, this tank was the largest ever built for storing cryogenic liquids (below −150°F). This container is double-walled, with an aluminum inner tank separated from an outer steel tank by 3 ft of perlite insulation. Its outer dimensions are 73 ft in diameter by 61.5 ft high. A photograph of the site at Lake Charles (Fig. 48) shows the storage tank and barge-mounted liquefaction unit.[18, 19]

While these tests were underway in Bayou Long, Constock concentrated on a crash program of research, development, and engineering analyses on all phases of the project, directed toward the commercial venture. Included were innovations in gas processing and liquefaction techniques, material evaluation and development, ship designs, cargo handling, storage tanks, etc. The most amazing aspect of the program was the way it was accomplished by its subsidiary, Constock Liquid Methane Corp. They operated with a skeleton staff directed by J. A. Murphy and housed in less than 500 square feet of space in a New York office building. Specific research and development assignments were doled out to the laboratories of the parent companies in Chicago and Ponca City, Okla. In addition, several consultants from universities were employed on a part-time basis, primarily to translate the research results into design criteria for practical applications. The bulk of the engineering design and construction was handled by J. F. Pritchard Co., Kansas City, Mo. (gas processing, liquefaction, and plant construction); Gamble Brothers, Inc., Louisville, Ky. (wood and insulation specialists); J. J. Henry Co., N.Y. (naval architects and marine engineers); and A. D. Little, Inc., Cambridge, Mass. (storage and cargo-handling methods).

By the spring of 1947, complete designs, specifications, and drawings for the liquefaction plant, tanker, and terminal facilities had been completed. Comprehensive analyses of potential gas sources and markets were also made. It was therefore possible to establish the economics for LNG shipments between a variety of ports all over the world. At this stage, a number of foreign countries, among which were England, Germany, France, Italy, Sweden, and Japan, expressed interest in importing LNG. By the fall of 1957, the British Gas Council made a declaration of intent to import LNG at a rate equivalent to 100 million cu ft of gas per day, which amounted to about 10 per cent of their total gas con-

sumption. However, signing a firm contract was deferred until after several trial shipments of LNG were made between the Gulf Coast and London. Even though the Constock ship designs appeared to be sound, there was as yet no proof that LNG could be transported overseas by tanker, particularly in rough weather

Constock agreed to erect liquefaction and land-storage facilities on the Calcasieu River near Lake Charles, La., while the British Gas Council supplied the unloading and terminal storage facilities at Canvey Island near London. Constock and the Gas Council agreed to share in the costs of converting a dry-cargo tanker, for this purpose, a new company, British Methane, Ltd , was formed to own and operate the ship

Lake Charles facilities The liquefaction plant, land storage, and loading terminal were erected on a twenty-acre site and placed in operation during the latter part of 1958 The barge-mounted liquefaction unit, which was used in the Bayou Long tests, was moved to Lake Charles A simplified flow sheet of the modified Claude (expander) cycle used on the barge is shown in Figure 47 Although this cycle is relatively inefficient, it has the advantage of being lighter and more compact, and therefore more adaptable to installation on a barge where space is limited It is interesting to note that this pilot liquefaction unit with a rated liquefaction capacity of 7 million cu ft of gas per day was 1 7 times larger in capacity than the Cleveland or Moscow plants of the 1940s

The liquid from the plant was stored in a tank having a capacity of 1 47 million gal (equivalent to 120 million cu ft of gas) Up until 1961, this tank was the largest ever built for storing cryogenic liquids (below −150°F) This container is double-walled, with an aluminum inner tank separated from an outer steel tank by 3 ft of perlite insulation Its outer dimensions are 73 ft in diameter by 61 5 ft high A photograph of the site at Lake Charles (Fig 48) shows the storage tank and barge-mounted liquefaction unit [18, 19]

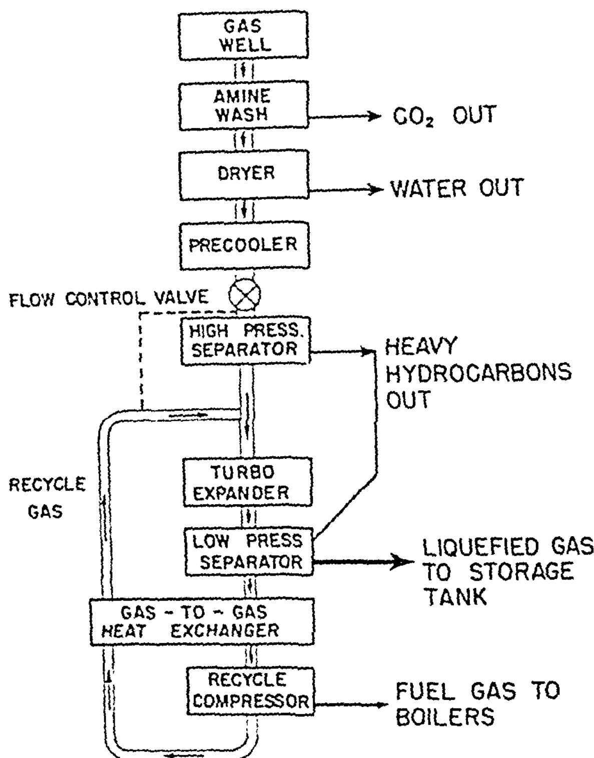


FIG. 47. Flow sheet of methane liquefaction barge.

Methane Pioneer: For transporting the LNG, a Cl-M-AVl dry-cargo ship (5,000-ton class) was converted to the MV *Methane Pioneer* at the Alabama Drydock & Shipbuilding Co. in Mobile, Ala., during 1958, according to plans and designs developed by Constock and J. J. Henry Co. A dry-cargo ship was selected because she has large double bottoms and wing tanks which can be used for ballasting, a particular problem raised by the low density of liquid methane (about 40 per cent of the density of water). In addition, other innovations in outfi[illegible] the ship for cryogenic service were re-

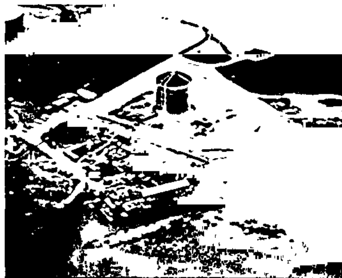


FIG. 48 Constock's LNG liquefaction terminal at Lake Charles, La

quired [19-21] However, only the insulation and cargo tanks will be mentioned here since they represent the major accomplishments upon which the success and economics of the project hinged [22]

The horizontal cross section of a tanker's hold space is essentially rectangular Therefore, in order to obtain the maximum utilization of this space for liquid cargo, the horizontal cross section of the cargo tanks must likewise be rectangular Cylinder tanks would be easier and much cheaper to fabricate, but unfortunately they would utilize only $\pi/4$ or about 80 per cent of the cargo space Because of the high cost of the ship—almost twice that of the conventional tanker—for LNG service, the economics dictate that the space utilization be greater than 90 per cent, thus prismatical tanks (rectangular parallelopipeds) are used Unfortunately, prismatical tanks introduce two severe design problems high stresses in the flat walls caused by cyclical, dynamic loads from rolling,

pitching, and heaving of the ship; thermal stresses in the walls caused by sharp, vertical temperature gradients when the tank is only partially filled. Obviously, one large tank filling the entire cargo space would be cheaper than several small ones. However, regulatory bodies for ships have limits on the size of individual compartments with respect to dynamic loadings, free-surface liquid effects, and safety under collision conditions.

Working within these shape and size limitations, the next step was to select the material of construction. Although all materials, metals, plastics, concrete, wood, etc. show an increase in strength with decreasing temperature, most of them become brittle at low temperatures. For this application, the choice narrows to aluminum (5,000 series alloy), stainless steel, or 9 per cent nickel steel.[23] Aluminum was selected on the basis of economics.

As was mentioned before, the prismatical tanks are expensive, so it was necessary to optimize the design. In order to do so, a method for making stress analyses had to be developed.[22] The general configuration of the tanks and relative location in the ship are shown in Figures 49 and 50.

In Figure 49, it can be noted that the aluminum tanks are surrounded with insulation, which had to meet the following requirements.

1. The insulation must maintain the ship's structure near ambient temperatures. Since the steel used in the hull cannot withstand LNG temperatures without becoming brittle, the insulation must also perform as a secondary, liquid-tight barrier in the event one of the aluminum tanks springs a leak.
2. The bottom insulation must be capable of withstanding the enormous stresses (due to the ship's motion) generated at the bottom key of the tank.
3. The insulation must provide a prescribed rate of boil-

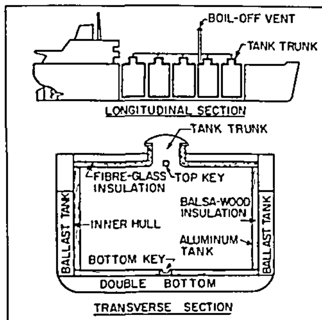


FIG. 49 Cross sections of vessel showing prismatical tanks

off of LNG if the vapors are to be used to generate power for propelling the ship

4 Since the insulation has to be attached to the ship's inner hull, it must be able to withstand the severe thermal stresses (−250°F on one face and ambient on the other) without yielding

5 In the event of fire on board ship, the insulation must be able to maintain structural integrity for at least four hours when its outer face is exposed to a temperature of 1200°F

Balsa wood was the only insulating material that was adequate to satisfy all these particular requirements In general, any material whose ratio of yield stress to thermal stress is greater than one can be subjected to cryogenic

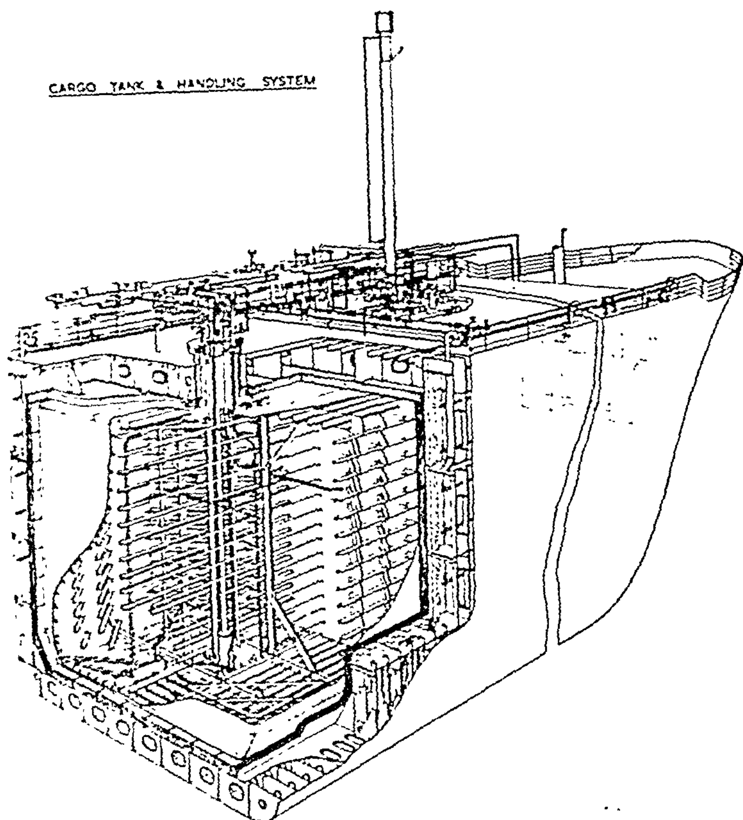


FIG. 50. Cutaway of cargo tank showing internal bracing.

temperatures in a fully restrained condition so that it is not free to contract. Table 13 compares this ratio for several common materials of construction. Although cast iron has a ratio of 2.3 it is not suitable for cryogenic service because it becomes brittle even at 0°F. Nine per cent *nickel steel* is not recommended for use below liquid nitrogen temperatures. With its high ratio, 2.2–9.2, wood is one of the best materials for cryogenic service.[4] Its chief disadvantage lies

4. Wood is even finding use at high temperatures, such as in spacecraft nose cones.[24]

Table 13. Ratio of Yield Stress to Thermal Stress for Common Materials

Material	$(S_{yp}/\alpha E(\Delta T)_{max}$*
Wood	2.2–9.2
Cast iron	2.3
9% Nickel steel	1.2
Foamglas	0.45
Concrete	0.3
Stainless steel (304 annealed)	0.28
Aluminum alloy (5,000 series, annealed)	0.22

*S_{yp} = Yield point in compression
α = Coefficient of thermal expansion
F = Modulus of elasticity
ΔT = Temperature differential, 70°F to −320°F = 390°F

in the difficulty to predict stresses. Being an anisotropic material, one must consider three ultimate strengths and 9 Poisson ratios in analyses.

The final design problem involves the insulation and cargo tanks together. Economics might indicate a thickness of insulation which results in prohibitive thermal stresses being generated in the tank walls, since the thermal stresses increase with decreasing insulation thickness, as shown in Figure 51.

The installation of the aluminum tanks is shown in Figure 52. Figure 52A shows the hold space lined with balsa panels which are so laminated as to give identical physical properties in two dimensions. The panels are faced with ⅛ inch plywood. They are 4 × 8 × 1 ft thick and are joined together with plywood scabs and fiberglass rosettes. Figure 52B shows one of the aluminum tanks, measuring 29 × 40 and 32 ft high, having a capacity of about 280,000 gal, being lowered into the hold. Figure 52C shows the first tank in place, the clearance between the sidewalls of the tank and insulation averages about 1 inch in order to maximize on cubic carrying capacity. Four other similar tanks, when installed, bring the total

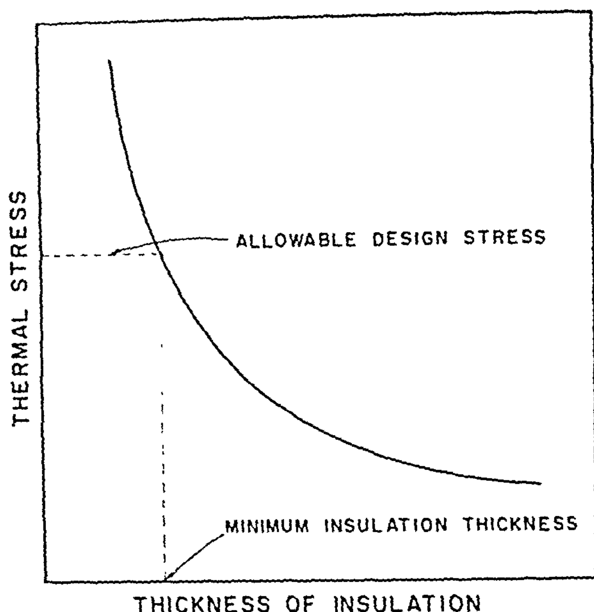


FIG. 51. Variation of thermal stresses in tanks with thickness of insulation.

capacity to over 1.40 million gal (equivalent to 115 million cu ft of gas).

Figure 53, left, shows a Cl-M-AV1 dry-cargo ship before conversion; right, the *Methane Pioneer* after conversion. The ship is over 350 ft long, 50 ft wide, and 40 ft deep, a healthy-sized pilot ship but tiny compared to present-day tankers. The *Methane Pioneer* took on her first load of LNG, equivalent to 115 million cu ft of gas, and departed from Lake Charles on her historic voyage, January 28, 1959. She arrived at Canvey Island on February 20, 1959, and discharged her cargo into two insulated, land-storage tanks, each having a capacity of 670,000 gal (equivalent to 55 million cu ft of gas).

Over the next year, ending in March 1960, the *Methane Pioneer* made six more trial shipments. The ship performed

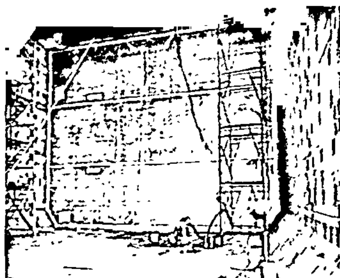
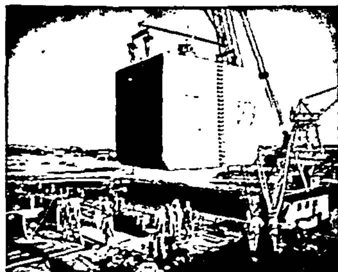
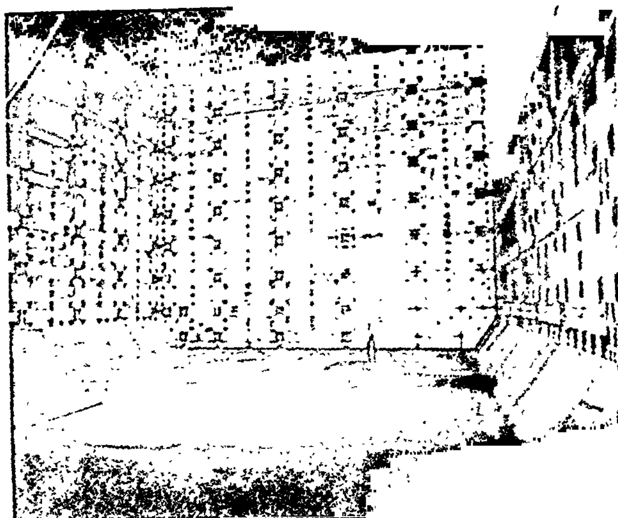


FIG. 52 Installation of aluminum tank in *Methane Pioneer*

A Insulated hold space



B Lifting the tank



C. Installed tank

admirably, even in very heavy seas where rolls exceeding 45° were recorded.[19, 20, 25] Having proved that LNG could be transported overseas by tanker, LNG shipments from Lake Charles were terminated because the *Methane Pioneer* was only a pilot ship and had less than one fifth of the minimum carrying capacity for economical operation. Since then, the *Methane Pioneer* continued in service as a refrigerated butadiene carrier between the Gulf Coast and Holland.

An idea which was conceived nine years earlier by W. W. Prince to solve a local power company's problem in Chicago had become an international enterprise. The world, particularly the skeptics—and there were many—awaited with great interest the maiden voyage of the *Methane Pioneer*. Hardly had she docked at Canvey Island before many others were dashing madly to get into the LNG business and reap the benefits at the expense of Constock's pioneering leadership. Constock, after nine years and some $15 million of investment had proved not only to herself but for all her potential



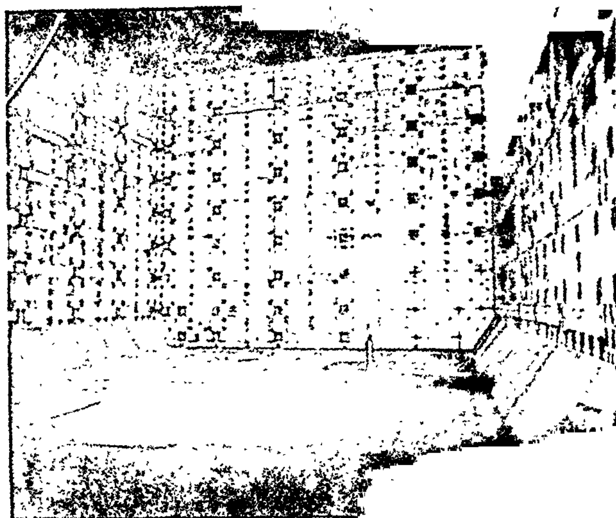
Fig. 53 Left, C-M-VI before conversion
Right, *Methane Pioneer* after conversion

competitors that the job could be done Although Constock had developed a protective wall of several hundred international patents, these alone were not enough to discourage competition

Commercial Ventures

Early in 1960, Royal Dutch/Shell joined forces with Constock, and a new company, Conch International Methane Ltd, was formed Royal Dutch/Shell and Continental Oil Co. each acquired a 40 per cent interest, and Union Stock Yards and Transit Co retained the balance of 20 per cent The headquarters were moved to London to initiate the first commercial venture for hauling LNG from North Africa to England The only activities that remained in the States were the pilot plant projects on gas purification, submerged pumping, in-ground storage, and fire tests, at Lake Charles, insulation development at Gamble Brothers in Louisville, and basic research at Continental Oil Co in Ponca City, Okla

The London headquarters were rapidly expanded to provide for research, development, engineering, marketing,



C. Installed tank

admirably, even in very heavy seas where rolls exceeding 45° were recorded.[19, 20, 25] Having proved that LNG could be transported overseas by tanker, LNG shipments from Lake Charles were terminated because the *Methane Pioneer* was only a pilot ship and had less than one fifth of the minimum carrying capacity for economical operation. Since then, the *Methane Pioneer* continued in service as a refrigerated butadiene carrier between the Gulf Coast and Holland.

An idea which was conceived nine years earlier by W. W. Prince to solve a local power company's problem in Chicago had become an international enterprise. The world, particularly the skeptics—and there were many—awaited with great interest the maiden voyage of the *Methane Pioneer*. Hardly had she docked at Canvey Island before many others were dashing madly to get into the LNG business and reap the benefits at the expense of Constock's pioneering leadership. Constock, after nine years and some $15 million of investment had proved not only to herself but for all her potential

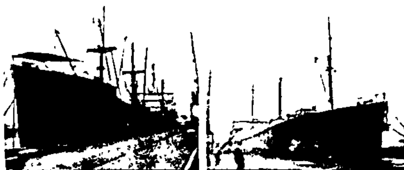


FIG. 53 Left, C-M-AV1 before conversion
Right, *Methane Pioneer* after conversion

competitors that the job could be done Although Constock had developed a protective wall of several hundred international patents, these alone were not enough to discourage competition.

Commercial Ventures

Early in 1960, Royal Dutch/Shell joined forces with Constock, and a new company, Conch International Methane Ltd, was formed Royal Dutch/Shell and Continental Oil Co. each acquired a 40 per cent interest, and Union Stock Yards and Transit Co retained the balance of 20 per cent The headquarters were moved to London to initiate the first commercial venture for hauling LNG from North Africa to England. The only activities that remained in the States were the pilot plant projects on gas purification, submerged pumping, in-ground storage, and fire tests, at Lake Charles, insulation development at Gamble Brothers in Louisville, and basic research at Continental Oil Co in Ponca City, Okla

The London headquarters were rapidly expanded to provide for research, development, engineering, marketing,

and patent services. Their first step was to initiate construction of a liquefaction plant and larger tankers modeled after the plans developed by Constock. Shortly thereafter, in the spring of 1960, Conch entered into an agreement with the French government to purchase gas[5] from the big Sahara Hassi R'Mel gas field and to build a liquefaction plant near Arzew, Algeria, on the North African Mediterranean coast. The liquefaction plant was to be financed and built by a new company, CAMEL (Cie Algierienne du Methane Liquid) owned jointly by Conch and French interests.

Although Britain presumably was to provide the initial market for this plant, and construction of the plant was undertaken, more than eighteen months elapsed before the U.K. government announced its approval in January 1962.[26] The delay was due to political angles involved: coal-industry opposition and the Algerian turmoil following its liberation, which raised questions regarding the stability of the Saharan gas as a source for imports. However, in the final analysis, the British Gas Council decided that the Sahara gas was as reliable as the British sources for oil which—in the minds of many Europeans since the Suez crisis—are not too certain.[27]

After Algeria gained its independence, it demanded and obtained additional participation in the LNG venture. The final lineup of operating companies is as follows:[28]

1. Liquefaction Plant by CAMEL (40 per cent Conch, 40 per cent French interests, and 20 per cent Algerian Government).
2. Gas Production at Hassi R'Mel Gas Field by Ste. d'Exploitation des Hydrocarbures d'Hassi R'Mel.
3. Pipeline[6] from gas field to Arzew by SOTHRA (combination of French and Algerian interests).

5. Contrary to popular belief, the contract price for the gas was around 25 cents per 1,000 cu ft, which is no better than current gas prices on our Gulf Coast

6. Three hundred miles of 24-inch pipeline.

4. Transportation by British Methane Ltd (50 per cent Conch and 50 per cent British Gas Council) which buys LNG, ships to England, and sells to British Gas Council.
5. Transportation by Gaz de France, the French counterpart of the British Gas Council, which buys LNG and ships to France [29]

With the decision by the French to participate in the initial venture by contracting for 50 million cu ft of gas daily, the capacity of the liquefaction plant was raised from 100 million to 150 million cu ft per day, with Britain under a fifteen-year contract to take 100 million cu ft per day. The liquefaction cycle selected was based on a cascade cycle developed by Constock. The terminology "cascade" derives its usage from the fact that the gas is progressively cooled in a series of steps, as shown schematically in Figure 54 The second law of thermodynamics provides that, as the difference in temperature between the refrigerating medium and the medium being cooled is diminished, the more efficient the process becomes In the limit, the most efficient process would be one in which the temperature difference was zero However, practical cost limitations force a compromise to a finite temperature difference, which as noted in Figure 54 amounts to 5°F in each step except one of the ethylene stages (8°F)

A comparison of the power and fuel consumptions for the various LNG liquefaction plants that had been built by 1965 is presented in Table 14. Although the theoretical, ideal power requirement is only 185 horsepower per million cu ft of gas per day, the practical fact that no machinery such as compressors and expansion engines have been built as yet which are 100 per cent efficient, and all economical heat exchangers must operate with finite temperature differences, raises the theoretical limit from 185 horsepower to a so-called practical limit (by today's standards) of 400 horsepower The Arzew plant is not far away from this limit with 469 horsepower

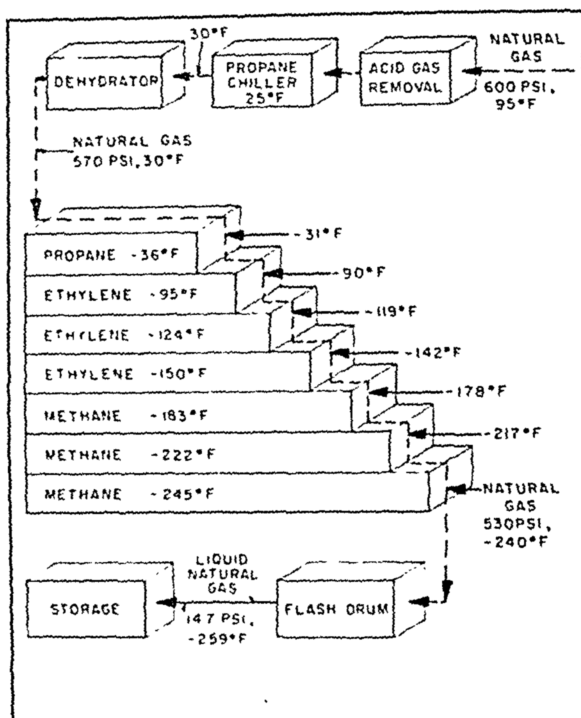


FIG. 54. Schematic of liquefaction of natural gas by means of cascade cycle in Algeria (data from *Chemical Engineering, 71,* 182. Oct. 12, 1964).

The storage facilities for liquiefied natural gas at Arzew amount to 17 million gallons of liquid or equivalent to 1 billion cu ft of gas.[28] There are three double-walled, aboveground storage tanks, each having a capacity of over 2.5 million gal. These tanks are modeled after the Lake Charles tank with a scaleup factor of about two. Two of these tanks have an inner shell of aluminum and the third a 9 per cent nickel steel inner tank. All three have 3 ft of perlite insulation between these inner shells and an outer shell of conventional steel.

TABLE 14. COMPARISON OF LNG LIQUEFACTION CYCLES

Cycle	*Inlet Gas Pressure, psi*	*Horsepower Per Million Cu Ft Gas/Day*	*Per Cent of Total Feed Used as Fuel*
Ideal	500	185*	
Practical limit	500	400†	
Cleveland cascade	615	662*	15
Russian cascade	725	660†	
Constock barge expander	1,000	1,000*	30
Arzew cascade	465	469*	8
Conch expander	465	540*	10
Transco cascade peak-shaving	315–490	716§	

*C. L. RITTER, *Chem. Engr. Progr.*, 58, 61–69 (1962)
†Estimated
‡P. S. PARKER and R. H. ARMSTON, *Gas Age*, *117* (March 11, 1954)
§R. MARTIN, *Petrol. Management*, *36*, 81 (Dec. 1961) (actual horsepower required is 624)

The rest of the storage, which amounts to 55 per cent of the total, or 9 million gal, will be in a frozen cavity 122 ft in diameter and 114 ft into the ground. This in-ground storage principle was developed by Conch[30] and is based on the principle that the moisture in the ground freezes to line a cavity, and serves as an ice barrier that prevents leakage and at the same time supplies adequate insulation. The use of this in-ground storage results in a saving over conventional above-ground storage in the neighborhood of about 5 cents per gal or, in this case, about $150,000.

The storage facilities will supply three tankers, each of them making about thirty round trips per year, and delivering about 7 million gal of liquid (equivalent to over 500 million cu ft of gas) per trip. Two of the tankers, the *Methane Princess* and the *Methane Progress* were built in British and Irish shipyards respectively. These tankers are essentially replicas of the *Methane Pioneer* in having prismatical aluminum tanks and balsa wood insulation. However, the French-

built tanker, appropriately christened the *Jules Verne*,[7] has seven cylindrical tanks of 9 per cent nickel steel insulated with 17 inches of Klegecell[8] on the bottom (conical-shaped, rather than keyed). The vertical walls and roof consist of 2½ inches of Klegecell covered with a liquid-tight sealant to act as a secondary barrier. The walls and roof are further insulated with perlite.[31] Apparently, the French were able to realize sufficient economies in going to the cylindrical, rather than prismatical tanks, despite the 20 per cent loss in cubic capacity. It has been reported[31] that the French ship cost $10 million to build against the $13 million for its British counterpart; the $3 million savings could have tipped the balance for cylindrical tanks, not to mention a possibility that the Conch-patented ship designs might have been evaded.

The terminal facilities on Canvey Island consist of seven conventional above-ground storage tanks, including the two that were used on the trial shipments with the *Methane Pioneer*, for a total capacity of 14 million gal or a gas equivalent of over 1.1 billion cu ft. The liquid is vaporized from storage in a novel scheme of heat exchange with propane and seawater which is capable of producing about 1.7 megawatts of power as a bonus. The vaporized gas then goes into an 18-inch pipeline extending some 300 miles northwest for delivery to 8 local gas boards, which will either blend the natural gas with manufactured gas or reform it to the lower calorific value needed in domestic burners.[32]

The investment costs for an operation liquefying and transporting LNG are enormous; an order of magnitude estimate is given in Table 15. To these figures must be added the cost of the unloading terminal facilities, which are comparable in cost to the liquefaction plant if gas reforming and pipeline distribution facilities are included.

7. French science fiction writer who half a century ago referred to natural gas as a major energy source.

8. Polyvinylchloride.

TABLE 15. APPROXIMATE COSTS FOR LIQUEFACTION AND OVERSEAS TRANSPORTATION

1. Capital investment			
a Liquefaction plant 100 million cu ft/day		40×10^6	
b. 2 Tankers @ 13×10^6			
Basic ship	50%		
Cargo handling	20		
Insulation	12		
Alum tanks	18	26×10^6	
			66×10^6
2 Annual operating costs (source to market about 1,500 miles)			13×10^6
3 Direct cost liquid natural gas delivered *before* taxes, administrative, research, and sales expense and *excluding cost of* gas at source		$0 36 1000 cu ft	

The costs given in Table 15 are to be viewed with reservation since they can vary widely depending on the locality For example, one reference[14] quotes the cost of the Arzew liquefaction plant as $60 million Even correcting the figure of $40 million given in Table 15 for a 100- to 150-million cubic feet per day plant, the corresponding figure would be around $50 million A later reference[32] pegs the total costs in Algeria at $70 million, another reports $86 8 million [28] From these figures it is possible to arrive at an estimated total investment in this first LNG project as summarized in Table 16 in round figures.

Whatever the final investment figures become, Conch is under a 15-year contract to supply Britain with 100 million cu ft of gas per day at a price of 88 cents per thousand cu ft, which indicates a slow payout The corresponding costs of manufactured gas in England from both oil and coal range from $1.07 to $1 75 per thousand cu ft depending on the process used.[27]

TABLE 16. REVISED ESTIMATE OF TOTAL INVESTMENT IN FIRST LNG VENTURE

		Millions of Dollars Invested
Gas pipeline (300 miles of 24-inch line from Sahara gas fields to Arzew)		$ 35.0
Liquefaction plant, Arzew		85.0
Two tankers		25.0
Unloading and distribution in London		
Canvey Island facilities	= $ 8.0	
Reforming plant	= 21.0	
Distribution line	= 21.0	50.0
Total		$195.0*

* This total does not include the French tanker and the unloading and distribution facilities in France. However, the gas pipeline in Africa was constructed apart from this particular project. Its cost probably balances the French investment in a tanker and unloading terminal. Therefore, the total investment resulting directly from the gas liquefaction is in the neighborhood of $200 million.

Future of LNG Transportation

The first commercial shipment of LNG from Algeria arrived in London on the *Methane Princess*, October 12, 1964. Since then, the *Methane Progress* and the *Jules Verne* have been brought into service. Much of the excitement regarding the potential markets for LNG in Western Europe was chilled to subzero proportions when gas was discovered on July 29, 1959,[33] by NAM (Nederlandse Aardolie Maatschapij—jointly owned by Royal Dutch/Shell and Standard Oil of New Jersey). The extent of the reserves was not publicized until around the middle of 1962, when it became rather evident that the field was the largest in Europe.[35] Up to this time, the Po Valley in northern Italy and Lacq in southeastern France provided the bulk of the natural-gas supply in Europe. By 1964, it became clear that the Dutch field was the third largest in

the world, ranking behind only the Texas Panhandle and the Sahara field. Although the extent of the Dutch discovery was not known prior to 1963, Royal Dutch/Shell as a partner to this discovery surely must have been aware of the ultimate impact of Dutch gas on LNG when they joined forces with Constock in 1960

It is certain that the Dutch gas will cut deeply into the formerly potential LNG markets in Holland, Belgium, Sweden, northern France, and Germany. The Dutch discovery has kicked off an exploration panic across the North Sea and extending inland into northern England. Engineering feasibility studies of laying a gas pipeline across the English channel are already in the advanced planning stages

France is attacking the natural gas supply problem on all fronts. The *Jules Verne* had not even made its first shipment when France announced plans to build another tanker four times as large.[31] For several years they had studied and experimented on a pipeline under the Mediterranean to bring Algerian gas to Spain and on into France,[36, 37] but recently admitted that the project was shelved because of Franco-Algerian political problems [31] However, at the same time, Ben Bella revealed his plans for a Mediterranean gas pipeline to Europe with a second gas liquefaction plant in Eastern Algeria.[38]

Even with the Dutch finally announcing that they were pricing their gas competitively with other fuels, as low as 35 cents per 1,000 cu ft, while continuing with expansion of distribution lines, Jersey Standard, the other half of NAM (discoverers of, and partners in producing Holland gas) gave their answer to the North Sea successes by releasing plans for shipping Libyan gas to Italy and Spain. The liquefaction plant, which will be located on the Mediterranean coast, 100 miles from the Essos Zelten area fields, will have a capacity of 300 million cu ft of gas per day — twice the size of the Arzew plant. Tankers will start shuttling to Italy and Barcelona,

TABLE 16. REVISED ESTIMATE OF TOTAL INVESTMENT IN FIRST LNG VENTURE

		Millions of Dollars Invested
Gas pipeline (300 miles of 24-inch line from Sahara gas fields to Arzew)		$ 35.0
Liquefaction plant, Arzew		85.0
Two tankers		25.0
Unloading and distribution in London		
Canvey Island facilities	= $ 8.0	
Reforming plant	= 21.0	
Distribution line	= 21.0	50.0
Total		$195.0*

* This total does not include the French tanker and the unloading and distribution facilities in France. However, the gas pipeline in Africa was constructed apart from this particular project. Its cost probably balances the French investment in a tanker and unloading terminal. Therefore, the total investment resulting directly from the gas liquefaction is in the neighborhood of $200 million.

Future of LNG Transportation

The first commercial shipment of LNG from Algeria arrived in London on the *Methane Princess,* October 12, 1964. Since then, the *Methane Progress* and the *Jules Verne* have been brought into service. Much of the excitement regarding the potential markets for LNG in Western Europe was chilled to subzero proportions when gas was discovered on July 29, 1959,[33] by NAM (Nederlandse Aardolie Maatschapij—jointly owned by Royal Dutch/Shell and Standard Oil of New Jersey). The extent of the reserves was not publicized until around the middle of 1962, when it became rather evident that the field was the largest in Europe.[35] Up to this time, the Po Valley in northern Italy and Lacq in southeastern France provided the bulk of the natural-gas supply in Europe. By 1964, it became clear that the Dutch field was the third largest in

the world, ranking behind only the Texas Panhandle and the Sahara field. Although the extent of the Dutch discovery was not known prior to 1963, Royal Dutch/Shell as a partner to this discovery surely must have been aware of the ultimate impact of Dutch gas on LNG when they joined forces with Constock in 1960.

It is certain that the Dutch gas will cut deeply into the formerly potential LNG markets in Holland, Belgium, Sweden, northern France, and Germany. The Dutch discovery has kicked off an exploration panic across the North Sea and extending inland into northern England Engineering feasibility studies of laying a gas pipeline across the English channel are already in the advanced planning stages.

France is attacking the natural gas supply problem on all fronts The *Jules Verne* had not even made its first shipment when France announced plans to build another tanker four times as large.[31] For several years they had studied and experimented on a pipeline under the Mediterranean to bring Algerian gas to Spain and on into France,[36, 37] but recently admitted that the project was shelved because of Franco-Algerian political problems.[31] However, at the same time, Ben Bella revealed his plans for a Mediterranean gas pipeline to Europe with a second gas liquefaction plant in Eastern Algeria [38]

Even with the Dutch finally announcing that they were pricing their gas competitively with other fuels, as low as 35 cents per 1,000 cu ft, while continuing with expansion of distribution lines, Jersey Standard, the other half of NAM (discoverers of, and partners in producing Holland gas) gave their answer to the North Sea successes by releasing plans for shipping Libyan gas to Italy and Spain. The liquefaction plant, which will be located on the Mediterranean coast, 100 miles from the Essos Zelten area fields, will have a capacity of 300 million cu ft of gas per day—twice the size of the Arzew plant Tankers will start shuttling to Italy and Barcelona,

with the former taking 80 per cent of the cargo, by late 1967. Total investment will be around $200 million.[39]

While others are looking at Middle East and West Pakistan gas for transportation to Japan, Australia, and South Africa, Polar LMG Corp., a subsidiary of Union Oil and Marathon Oil Co., is negotiating a contract with Tokyo Gas. Co. to deliver 35 million cu ft of gas per day from either Alaska or British Columbia.[40, 41] Originally, the timetable looked for 1965 as a shipping date, but it was delayed to 1967 after natural gas was discovered in Niigata, Japan.[42]

Scandinavian oil industry leaders are talking about a joint venture to import Sahara gas, despite the proximity of the Holland and North Sea fields.[42]

It is no secret that Conch, Union Oil, Ohio Oil, and others are looking at both the east and west coasts of the United States with an expectative eye. California's demand for gas is expected to reach 8 billion cu ft of gas per day by 1982, at which time it is estimated that they will be producing less than 1 billion cu ft.[43, 44]

Australia, which is still considered a good market, is continuing to show substantial increases in their own gas fields.[45]

Venezuela flares more gas by a factor of almost two than is currently consumed in all of Western Europe. They missed their chance of becoming the first source for LNG when they could not come to an agreement with Constock on gas price back in 1958,[14] but they very likely will enter the picture soon.

Looking to the year 2000, when the world's energy consumption will climb to five times the current rate,[1] it is reasonable to conclude that transportation of LNG will continue to prosper. Even though natural gas continues to be discovered all over the world, there is an interesting opposing force which will act to increase the cost burdens on distribution by pipeline. As population density increases, right-of-way cost for pipelines rises drastically. Holland, with a population density of 910 per square mile, coupled with the need for

myriads of underground road and water crossings, is faced with pipeline construction costs that are two to three times higher than the average elsewhere.[34] Already they are resigned to the fact that at least half their production will have to be exported

Competition from other fuels will always be a problem, but the petroleum and gas industries thus far have demonstrated a unique—so far as energy sources are concerned—adaptive, pricing-control characteristic which keeps its products moving on the open market. Continued research and development will lower the cost of producing LNG Modest improvements can be expected in the liquefaction and storage aspects; however, the next big jump will occur when shipbuilders abandon the notion that an LNG carrier must look and act exactly like a conventional oil tanker and start building a carrier to conform to the needs of the cargo Some trends have already been noted, ranging from modest[46-50] to others more daring [51-53]

II PEAK LOAD SHAVING AND OTHER USES

At least one promising use for LNG which does not involve the liquid transportation problem is known as peak-load shaving, or simply, peak shaving This concept can be most easily described by referring to Figure 55, in which the variation in daily demand for gas by a local utility throughout the year is shown. It is obvious that the sinusoidal curve represents a highly idealized smoothing of the actual curve, which has numerous peaks and crevices The demand peaks on the coldest day in winter and bottoms on the hottest day in summer. The yearly average daily demand is represented by the horizontal broken line, and the shaded areas labeled deficit and surplus are equal In order to obtain the best price and to be guaranteed a supply, the local utility must contract purchase from the gas pipeline company at a fixed rate da

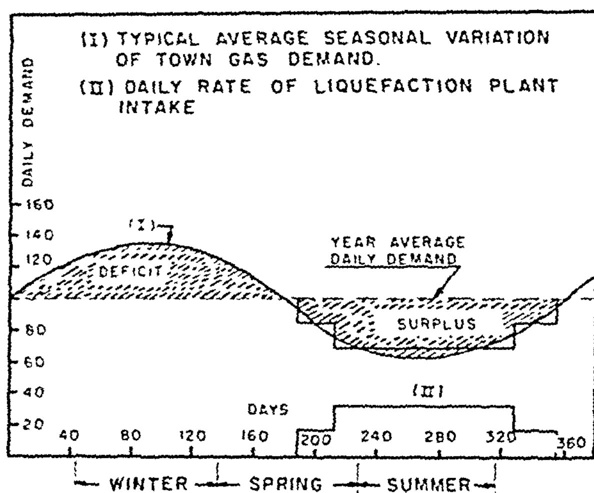


FIG. 55. Variation in daily gas demand with seasons of the year.

year round. If the purchases are made on the yearly average daily rate, customer demand in the winter months cannot be met. On the other hand, there will be a surplus of gas available in the summer months.[9] The obvious solution is to "save" the *surplus* in the summertime and draw on it to make up the deficit in the wintertime. However, because huge volumes of gas would have to be stored, no man-made storage is practical.

In some regions of the country there are depleted gas and/or oil fields or aquifer[10] where large volumes of gas can be stored underground. In this case the solution is simple; the utility company stores surplus gas in the natural underground reservoir and draws on it during cold weather when consumer demands exceed pipeline capacity. The amount of

9. The treatment here is considerably simplified. Gas distribution and pricing is a complex matter.[58]

10. Aquifer here refers to a geologic anticline or trap having a stratum of permeable water-bearing sandstone or limestone which can be developed for storage of gas. In some cases, salt caverns or abandoned coal mines can be used for storing gas.

such storage in the United States is increasing rapidly; in 1962, the capacity was 3.56 trillion cu ft (about half is usable) as compared to a total marketed gas production of 13.3 trillion cu ft.[55] The big storage concentration, fortunately, is in the more populated areas in such states as Pennsylvania, Michigan, Ohio, West Virginia, California, and Illinois. Nevertheless, there are many areas, particularly New England, the East, and the South, where other measures have to be taken. For example, the surplus gas could be sold to local industry at a bargain (interruptible) rate, in which case synthetic gas would have to be used to supply the deficit. The synthetic gas is frequently manufactured by thermal cracking of residual oils, catalytic reforming of propane and butane, or blending propane and air with the pipeline gas.

As an alternative to manufactured gas for peak loads, but not competitive with natural underground gas storage, natural gas could be liquefied and stored during the summer months and then vaporized during the winter months to augment the pipeline supply to meet peak demands. The Cleveland plant was built for this purpose in the early 1940s. Within the last ten years, technological improvements in liquefaction equipment and cryogenic storage have tipped the economic balance in favor of LNG over manufactured gas for peak-shaving. Realizing this potential, Constock International Methane Ltd and J. F. Pritchard Co. formed a new company, Constock-Pritchard, Inc., early in 1958, to develop and construct LNG peak-shaving plants.

As a further boon to LNG peak-shaving, the concept of in-ground storage for LNG was developed in the early 1960s. Working independently, Conch,[39, 56, 57] the American Gas Association,[58, 59, 5] and Phillips Petroleum Co.[60] each developed their own versions. A schematic of the in-ground storage, piloted by Conch at Lake Charles in 1961, is shown in Figure 56. The construction of the cavity-in-ground (CIG) is as follows:

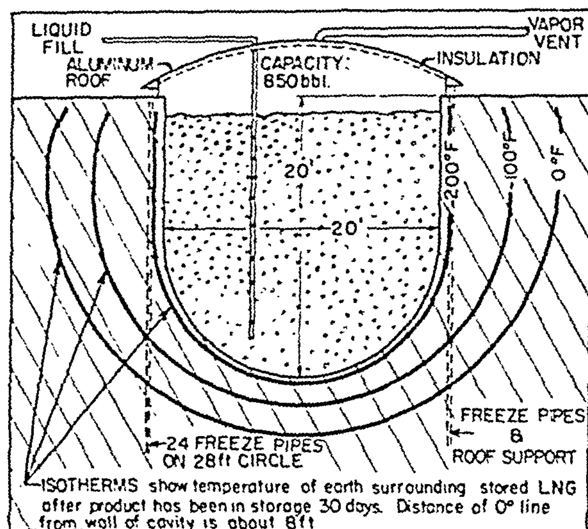


FIG. 56. How liquid methane is stored in frozen ground by Conch.

1. Freeze pies for circulating refrigerant are sunk to the required depth along the periphery of a circle whose diameter is 20 to 40 per cent greater than the diameter of the intended cavity.
2. The ground is prefrozen by circulating an external refrigerant, such as propane, to form a frozen ring.
3. When the inner diameter of the frozen ring reaches the diameter of the cavity, the unfrozen ground, bounded by the frozen ring, is excavated.
4. During excavation, an insulated roof is installed.
5. The cavity is now ready for filling with LNG.

The prefreezing operation eliminates the need for costly shoring during excavation since the frozen ground has structural strength comparable to concrete. It also prevents moisture penetration during excavation. The frozen ground acts as an impermeable barrier to any penetration or loss of LNG

into the surrounding ground and also provides adequate insulation to prevent excessive boil-off during storage

The temperature distribution measured during the Lake Charles tests is also shown in Figure 56. The problem of theoretically predicting these isotherms and the rate of heat leakage into the cavity is not easy.[5, 61-63] It is necessary to know the heat leak quite accurately in order to establish the size of the refrigeration equipment needed.

Equally important is a knowledge of the properties of frozen ground, concrete, and other construction materials at these low temperatures in order to be able to predict the structural integrity of the cavity with time [23, 64-66] Review of the design problems involved in cavity-in-ground storage is given by Sharp[67] and by Khan et al [68]

A number of articles on the economics of peak shaving with LNG have been published.[7, 68-70] The basic differences between LNG transportation and LNG peak-shaving plants are

1. Peak-shaving plants tend to be of much smaller liquefaction capacity, but their vaporization capacity is larger since they must be able to deliver gas at a very high rate for the few (5–15) coldest days in the year
2. The liquefaction section of peak-shaving plants lies idle one third to one half the year, whereas in LNG transportation it operates around the clock
3. Peak-shaving plants require relatively more storage than LNG transportation plants The cost of storage in the former runs over 25 per cent of the total investment, whereas in the latter it is usually below 10 per cent
4. Even though the peak-shaving plants require no investment in tankers, loading and unloading docks, reforming facilities, etc , the total investment per unit of liquefaction capacity tends to run 10–20 per cent higher than LNG transportation plants Both of them run pretty close to $1 of total investment per cu ft of gas liquefied

5. The unit selling price of gas delivered to the pipeline will run about 30–50 per cent higher in a peak-shaving plant than an LNG transportation plant, depending on whether reforming costs are included in the latter. Round figures for the LNG transportation plant are $1.10 per 1,000 cu ft as compared to $1.50 per 1,000 cu ft for peak-shaving.

It is to be understood that the above figures are used for illustrative purposes only, in an attempt to typify and compare a peak-shaving plant having a liquefaction capacity of around 10 million cu ft of gas per day with a transportation plant of 100 million cu ft of gas per day. Depending on the location of the site, these figures can vary drastically.

Four LNG peak-shaving plants in this country were scheduled to begin operations in 1965. The first one went on stream early in 1965. It was built by Constock-Pritchard Inc. for Transcontinental Gas Pipeline Corp. (Transco) on 420 acres in the Jersey meadows near South Hackensack, N.J., at a reported cost of $12 million. It will have a liquefaction capacity of 5 million cu ft per day into 1 billion cu ft of equivalent gas storage with a maximum vaporization capacity of 200 million cu ft of gas. Transco will initially sell the gas to eight utility companies from Georgia to New Jersey for winter peak-shaving at a cost (reported to be $2.20 per 1,000 cu ft) about half to three quarters the cost a utility would pay to manufacture its own peak-shaving gas. Because of the marshy nature of the ground, which rendered above-ground storage impractical, Transco will use in-ground storage consisting of a cavity 115 ft in diameter and 165 ft deep,[71] with a capacity of 12 million gal of liquid or a billion cubic feet of gas.

The second LNG peak-shaving facility is under construction at a cost of $3 million near Birmingham, Ala.; it will be jointly owned by Air Products and Chemicals Co. and Alabama Gas Co. This plant is slightly smaller than Transco's.

It has a daily liquefaction capacity of about 4 million cu ft of gas and a vaporization capacity of 85 million cu ft. The storage consists of a conventional, above-ground, double-walled tank having an inner liner of 9 per cent nickel steel and insulated with 5 ft of perilite. The storage capacity is 7.2 million gal or 600 million cu ft of gas [72] In proximity to this plant, Air Products is involved in a cryogenic complex which includes a gas-reforming plant to produce hydrogen from natural gas. The hydrogen is used for manufacturing ammonia and as feed to a 30 ton per day hydrogen liquefaction plant which will supply NASA's Pearl River test site near New Orleans. In addition, some of the LNG from the peak-shaving plant might find outlet as a rocket fuel. A multiplicity of cryogenic products from a single complex offers intriguing economic returns and could set the stage for a healthy growth in cryogenic processing, which, to date, has been monopolized by the production of liquid nitrogen and oxygen.

A third LNG peak-shaving plant was planned for late 1965 at Chula Vista, Calif., built by American Messer Corp for San Diego Gas and Electric Co The liquefaction plant will use pipeline pressure drop to liquefy natural gas in an expander cycle at a capacity of 2 million cu ft of gas per day. The storage tank is identical with the one at Birmingham except that it uses 3 ft of perlite insulation instead of 5 ft. The vaporization capacity is 60 million cu ft of gas per day The cost is $2.7 million [73]

A fourth peak-shaving plant was built for Wisconsin Natural Gas Co at Oak Creek, Wis., between Racine and Milwaukee. The plant has an above-ground metal, double-wall tank providing the equivalent of 250 million cu ft of gas storage, a vaporization capacity of 50 million cu ft of gas per day, and a net liquefaction rate of 750,000 cu ft of gas per day. Chicago Bridge and Iron Co are the designers and builders of the plant, which was scheduled to go on stream in the fall of 1965 [74]

As LNG peak-shaving plants become more widespread across the United States, other uses for LNG will develop because of its increased general availability at reduced costs. Two promising possibilities are as a rocket propellant and as a fuel for supersonic aircraft.

Prior to 1959, LNG was considered a possible candidate as a rocket fuel. However, it seemed to drop out of the picture when NASA took over the space program and concentrated on such fuels as RP-1,[11] hydrazine, and liquid hydrogen. Within the last several years, NASA has had to modify its thinking somewhat. Although liquid hydrogen, in combination with liquid oxygen, provided the high performance characteristics needed, there were some applications, such as space storability, recoverable vehicles, etc., for which the hydrogen-oxygen system—as well as other fuels—was unsuited. In this respect, LNG offers the most attractive alternative. When LNG is used in combination with the oxidizer, FLOX (a liquid mixture of fluorine and oxygen), its performance is commensurate to the hydrogen-oxygen system.[12] Compared to RP-1, hydrazine, and hydrogen it is the safest material to handle,[75] and its cost is substantially lower than hydrazine and hydrogen, being only slightly higher than RP-1. In addition, its properties are such that it can be adapted to existing rocket hardware, or it can lead to simplification in new hardware.

The enormous heat dissipation problem presented by supersonic jets will require the development of more thermally stable fuels which can serve as a heat sink for cooling the engine and the leading edges of the air frame, in addition

11. Similar to jet fuel.

12. Another interesting comparison of performance is afforded by the 3-stage *Saturn V* rocket. If the present RP-1/LOX booster stage were replaced by LNG/FLOX, the payload for solar system escape could be increased from 3,500 to 10,300 lb. This performance even exceeds the payload of 7,000 lb which could be put into solar system escape by replacing the third stage, LH_2/LOX, in the present *Saturn V*, with a nuclear engine.

to maintaining the required temperature in the cabin for both the payload and electronic instrumentation. Some of the thermally stable fuels that have been considered cost twenty to fifty times as much as jet fuel. Whether or not supersonic jets will ever become a reality hinges on fuel economy. LNG, with its superior performance and cooling capabilities, added to the fact that its cost is comparable to jet fuel, offers the most promising solution.

Both of these applications are being investigated by the Research Division of Continental Oil Co For their studies, and to supply test quantities for others, they have erected a liquefaction and purification unit at Ponca City that can produce up to 600 gal per day of liquid methane having a purity of 99.9+ per cent

The fact that LNG is currently not generally available should not serve as a deterrent to its ultimate use as a rocket or supersonic jet fuel

Even if LNG were not being developed for heating purposes, it would merit development for rocket and aircraft fuels Although the total consumption of LNG by the rocket industry would represent an insignificant fraction of the total natural gas currently produced, it can be readily liquefied at any site near a pipeline or imported as LNG at a cost substantially lower than the other rocket fuels, with the exception of RP-1 (and ammonia if it enters the picture). On the other hand, supersonic jets, with their voracious appetites for fuel, could eventually consume up to 10 per cent of the total natural gas production in the United States In this case, the demand is sufficient to realize the savings in manufacturing costs in large-scale plants, consequently, the LNG could be produced competitively solely for the purpose of replacing present-day fuels On the other hand, the day is rapidly approaching when LNG will be available in practically every major population center in the world to satisfy either a peak-shaving or a base-load demand Since supersonic jets

will largely confine their operations to such centers, the general availability of LNG *may* well exceed the current, general availability of jet fuel. Even today, certain preferred jet fuels are difficult to obtain abroad.

Summary

The next decade will probably reveal an unprecedented growth of a single commodity, liquefied natural gas, as a major factor in supplying the world's demands for energy. Although the pioneering efforts of Constock were primarily directed toward the foreign markets, the demand for LNG in the United States, both as peak-shaving and base load, could even surpass many of the foreign markets. Both the East and West coasts are feeling the pinch for more gas. Although the nation's 1,500 gas transmission and distribution companies are laying pipelines at the rate of about 30,000 miles per year and a cost of about $1.7 billion per year to an already established piping network of over 700,000 miles,[76] it is questionable whether they can keep pace with the growth in demand, notwithstanding predictions to the contrary.[77] The cost of constructing pipelines in this country varies from $10,000 to $280,000 per mile depending on the terrain and pipe size. An average cost for good pipelining country is around $100,000 for a 30-inch pipe, or roughly an average of from $2,000 to $3,000 per mile per inch diameter of pipe. These costs include right-of-way, surveying, communications for remote operation, maintenance, meter stations, miscellaneous materials, and compressor stations.[78] The cost of transporting natural gas by pipeline is about 1.7 cents per 1,000 cu ft per 100 miles, which is about double the corresponding cost for transporting it by LNG tanker. When one adds to the LNG figure the costs of liquefaction and regasification, the pipeline and LNG begin to approach an even trade-off for the same distance between the source and the

market In fact, Venezuelan gas for the East Coast and Alaskan gas for the West Coast, or possibly Mexican gas for either coast, could put the squeeze on pipeline gas, which is experiencing a steady rise in cost, year after year. If LNG does not dent the pipeline market, it certainly will establish a ceiling for the "city-gate" price. Furthermore, if gas prices at the producing fields continue to rise at present rates, coal shipped from the north by barge or tanker could displace gas for power generation in the southwest.

In the last analysis the total energy picture will depend greatly on the continued development of atomic energy and direct conversion devices such as the fuel cell, the thermionic converter, the thermoelectric generator, and the magnetohydrodynamic converter. Of these, the fuel cell and the magnetohydrodynamic generator will probably have the greatest effect on the increasing use of natural gas for generating power.

The use of LNG as a starting raw material for petrochemical processing represents a good, profitable usage even though the quantity, as compared to fuel consumption, is only a matter of a few per cent

Finally, political factors such as import regulations and nationalization of privately owned plants in foreign countries, could be the tail that wags the donkey

In retrospect, if the writer has failed to convey the significance of Constock's successful pioneering effort, then let it be remembered as an outstanding technical achievement In this era of daily breakthroughs, however, it is appropriate to quote Dr Harvey Brooks, Dean of Engineering and Applied Physics at Harvard University

> Breakthroughs are mainly figments of publicity agents' imagination They come slowly, a series of steps, advancing the art There's more prior art than we're led to believe It's a matter of sudden realization

The author appreciates the permission granted by Conch International Methane Ltd., Constock-Pritchard Corp., and Continental Oil Co. to use information obtained from these sources in the presentation of this paper.

REFERENCES

1. S. H. SCHURR, Energy, *Sci. American*, *209* 111 (1963).
2. P. SPORN, *Energy—Its Production, Conversion and Use in the Service of Man*, Macmillan (1963).
3. M. A. ELLIOTT, What's Ahead in Energy, *Petrol Management, 36,* 70 (1964).
4. Special Report, "Energy," *Power, 101,* C-40 (1957). See also J. H. LICHTBLAU and D. P. SPRIGGS, "Energy Policy and Competition," Project 12, Petroleum Industry Research Foundation, New York (1961).
5. B. E. EAKIN and C. G. VON FREDERSDORFF, Below Ground Storage of Liquefied Natural Gas, *Chem. Eng. Progr.*, *58*, 46 (1962).
6. C. V. SPANGLER, Stock Pile Natural Gas, *Oil and Gas J.*, *48*, 94 (1949). See also *Oil and Gas J., 49,* 170 (1950); and *Gas, 26,* 62 (1950).
7. A. R. YOUNG, "Recent Developments in Liquefaction of Natural Gas for Peak Saving," presented to the American Gas Ass. Production Conf., Pittsburgh, Pa. (May 17, 1963).
8. H. P. CADY, Beginnings of the Helium Industry, *Ind. and Eng. Chem.*, *30*, 845 (1938).
9. C. W. SEIBEL, Production of Helium at Amarillo, *Ind. and Eng. Chem. 30*, 848 (1938).
10. R. H. ORMSTON, Liquefaction: The Answer to Storage? *Gas Age, 117*, 27 (1954).
11. *Liquefied Natural Gas, A Bibliography*, American Gas Ass., New York (March 14, 1960).
12. J. BURNS and L. J. CLARK, *Liquid Methane*, Publ. 484 of the Institution of Gas Engineers, London (May 1956).
13. M. A. ELLIOTT, C. W. SEIBEL, F. W. BROWN, R. T. ARTZ, and L. B. BERGER, "Report on the Investigation of the Fire at the Liquefaction, Storage, and Regasification Plant of the East Ohio Gas Co., Cleveland, Ohio, October 20, 1944," U.S. Bureau of Mines Report, R. I. 3867 (Feb. 1946).

14 J. D. BALCOMB, *Liquefied Natural Gas and the International Energy Market*, International Energy Reports on United Nations Economic Developments, Worldmark Economic Publications, Inc., New York (1958).
15 Liquefaction Behind the Iron Curtain, *Gas Age, 117*, 29 (1954).
16 R. L. Huntington, *Natural Gas and Natural Gasoline Manufacturing*, McGraw-Hill, New York (1950), p 14.
17. Technology Frozen Gas, *Time Mag*, p 88 (Nov. 17, 1961)
18 J W. HUNT, "Methane Liquefaction and Storage Operations at Lake Charles, Louisiana," presented before the New England Gas Ass, Worcester, Mass (June 20, 1961)
19. C L. RITTER, Recent Developments in Liquefaction and Transportation of Natural Gas, *Chem Eng. Progr*, *58*, 61 (1962).
20 C. G FILSTEAD and M BANISTER, "Low Temperature, Liquefied Gas Transportation," presentation before The Society of Naval Architects and Marine Engineers Annual Meeting, New York (Nov. 16-17, 1961)
21. E M. SCHLUMBERGER and J W HUNT, "Design and Transportation Aspects in the Handling of Liquid Methane," presented at the Annual Tanker Conference of the American Petroleum Institute at Cape Cod (June 12-14 1961)
22. C M. SLIEPCEVICH, Materials and Process Design Calculations, *Chem Eng Education* p 15 (March 1964)
23. L P ZICK and M. B CLAPP, How to Specify Low Temperature Storage Vessels, *Petrol Refiner, 43*, 125 (1964)
24 D. G. HOLLMAN, Wood for Spacecraft Nose Cones, *Space/Aeronautics, 40*, 97 (1963)
25 Shipper Reveals How Methane Tanker Performed, *Chem Eng*, *68*, 64 (1961)
26 U K Stamps O K on Gas Lift From Africa, *Chem Eng*, *69*, 46 (1962).
27 Commons Urges Britain to Import Saharan Gas, *Oil and Gas J*, *59*, 81 (1961)
28 First Algerian Gas Enroute to Britain, *Oil and Gas J*, *62*, 105 (1964).
29. Subzero Shipping Opens New Era for Gases, *Chem Eng*, *69*, 64 (1962).
30. C. M SLIEPCEVICH, Ground Reservoir for the Storage of Liquefied Gases, U S. Pat 3,159,006 (Dec 1, 1964)
31 French Launch LNG Tanker, *Oil and Gas J*, *62*, 100 (1964)
32 British Gas Network Awaits African Methane, *Chem Eng*, *71*, 96 (1964)

33. Algeria Dreams of Petrochemical Future, *Chem. Eng. News*, *41*, 130 (1963).
34. J. H. Winchester, Holland Strikes It Rich in Natural Gas, *Readers Digest*, p. 83 (Aug. 1964).
35. A Look at the Common Market in the 1970's, *Oil and Gas J.*, *60*, 68 (1962).
36. O. Wolff, Pipeline Across the Mediterranean, *Oil and Gas J.*, *59*, 91 (1961).
37. French Ponder Sahara–Europe Gas Line, *Oil and Gas J.*, *61*, 92 (1963).
38. Ben Bella Has Big List of Oil Plans, *Oil and Gas J.*, *62*, 124 (1964).
39. Jersey Plans to Ship Libyan Gas to Italy and Spain, *Oil and Gas J.*, *62*, 79 (1964).
40. Union, Marathon Consider Alaskan Gas Liquefaction, *Hydrocarbon Process. and Petrol. Refiner*, *42*, 129 (1963).
41. Japan May Soon Use U.S. or Canadian Gas, *Oil and Gas J.*, *62*, 64 (1964).
42. HPI Newsletter, *Hydrocarbon Process. and Petrol. Refiner*, *42*, 12 (1963).
43. Liquid Methane Considered for West Coast Use, *Oil and Gas J.*, *59*, 61 (1961).
44. Huge Growth Seen for Liquid Methane Both Here and Abroad, *Oil and Gas J.*, *57*, 82 (1959).
45. Coming Next for Australia: Natural Gas Pipelines, *Oil and Gas J.*, *61*, 68 (1963).
46. Processing Notes, *Oil and Gas J.*, *61*, 140 (1963).
47. P. Amauric, Overseas Transportation of Liquefied Gas, *Compressed Air Mag.*, *68*, 16 (1963).
48. Expanding Tanks Raise Tanker Capacity, *Oil and Gas J.*, *61*, 38 (1963).
49. E. Schlumberger, "Recent Developments in Transportation, Liquefaction, and Safe Operating Practices for Natural Gas," presented at the American Gas Ass. Production Conference, New York (May 26, 1964).
50. Paul Endacott, A Unique Liquefied Petroleum Gas Carrier, *Shipbuilding and Shipping Record*, *104*, 11 (1964).
51. New Tanker Design Cuts Crew Two-Thirds, *Oil and Gas J.*, *61*, 64 (1963).

52. Almost Here: Tanker That Travels Underwater, *Oil and Gas J.*, *61*, 90 (1963)
53. Ocean Transportation of Liquefied Natural and Petroleum Gases, *Oil and Gas J.*, *57*, 76 (1959).
54 J A SHERRED, Who Will Pay Top Price for Natural Gas—Chemicals or Fuels? *Petrol. Refiner*, *36*, 101 (1957)
55 R. B BIZAL, Gas-storage Capacity Spurts, *Oil and Gas J*, *60*, 125 (1962)
56 Mudpie Provides New Method for Storing Liquefied Gases. *World Petrol*, *32*, 48 (1961)
57. H C BOZEMAN, Frozen Holes Provide LNG Storage, *Oil and Gas J*, *59*, 84 (1961)
58 Buried Concrete Tanks New LNG Storage Plan May Cut Costs, *Oil and Gas J*, *59*, 50 (1961)
59. Concrete LNG Tank Cuts Storage Costs, *Oil and Gas J*, *61*, 51 (1963)
60 Propane to Be Stored in Frozen Earth, *Chem Eng News*, *40*, 52 (1962).
61 D E FLANAGAN and P B CRAWFORD, Feasibility of Underground Storage of Liquefied Methane, *J Petrol Technol*, *12*, 73 (1960) See also *Oilweek*, *13*, 26 (1962)
62. S W. CHURCHILL, Heat Transfer Rates and Temperature Fields for Underground Storage Tanks, *Soc Petrol Eng J*, *225*, 28 (1962)
63 H HASHEMI, "Heat Conduction with Change of Phase," Ph D Thesis, Univ of Oklahoma, Norman (1963)
64. J O THIEME and R L EVERY, 'Thermal Properties of Clay Silt and Ice from Ambient to Cryogenic Temperatures," pre- [illegible]
65 G E MONFORE and A E LENTZ, Physical Properties of Concrete at Very Low Temperatures, *J Portland Cement Res and Dev Labs*, *4*, 33 (1962)
66 M SABBAGHIAN, 'Thermoviscoelasticity With Time and Stress Dependent Coefficient of Expansion' Ph D Thesis, Univ of Oklahoma, Norman (1964)
67 H R SHARP, Refrigerated Storage Requires New Techniques in Design, *Oil and Gas J*, *62*, 52 (1964)
68 A R KHAN, T J JOYCE, and J HUEBLER, "Status of LNG Storage," presented at the Production Conference of the American Gas Ass, New York (May 25-26 1946)

69. A. Russell Young, Liquid Methane—A Cheaper Means of Peak Shaving, *Oil and Gas J.*, *57*, 75 (1959).

70. A. W. Mellen, "The Economics of Peak Shaving in the U.S. Northeast," presented at the Annual Meeting of the American Institute of Chemical Engineers, Denver, Colo., Liquefied Natural Gas Symposium (Aug. 28, 1962).

71. R. Martin, First Cryogenic In-Ground Gas Storage in USA, *Petrol. Management*, *36*, 84 (1964). See also *Oil and Gas J.*, *62*, 92 (1964); ibid., *62*, 101 (1964); *61*, 70 (1963); *60*, 61 (1962).

72. Cryogenic Techniques Find New Roles in Natural-Gas Processing, *Oil and Gas J.*, *62*, 121 (1964); ibid., *62*, 48 (1964); and *Cryogenic Information Report*, *1*, Item III-9 (1964).

73. Liquefaction Plant to Use New Method, *Oil and Gas J.*, *62*, 27 (1964).

74. H. E. Vaughn, Liquefied Natural Gas Projects Today, *Petrol. Refiner*, *44*, 131 (1965).

75. *Liquefied Natural Gas—Characteristics and Burning Behavior*, Conch Methane Services Ltd., Univ. of Oklahoma Research Inst., Norman (1962).

76. AGA Predicts 4% Gain for Gas in '64, *Oil and Gas J.*, *61*, 80 (1963). See also ibid., *61*, 68 (1963).

77. Liquid Methane No Present Threat to U. S. Gas Lines, *Oil and Gas J.*, *57*, 113 (1959).

78. Seventh Annual Study of Pipeline Installation and Equipment Costs, *Oil and Gas J.*, *62*, 87 (1964).

11. SYSTEMS ANALYSIS AND THE VISUAL ORIENTATION OF ANIMALS

By TALBOT H. WATERMAN
Yale University

The purpose of this paper is twofold Most directly it reviews our current research on the visual responses of aquatic animals, mainly crustaceans In addition, it attempts to demonstrate the usefulness or perhaps the necessity of using some appropriate formalism like systems analysis in the adequate study of biological problems even of a relatively simple sort.

The need for an explicit methodology of this kind arises from the nature of biology Living beings, and even their component organs, tissues, and cells, are multivariable systems *in which numerous elements are complexly interrelated* As a result, few of these variables are purely dependent or purely independent Furthermore, despite spectacular progress in some areas of the life sciences, many essential components and their interrelations remain quite unknown Consequently, some formal strategy is clearly necessary in biology for effectively formulating questions, planning experiments, and analyzing data

Systems analysis, or similarly structured general analytical approaches like control theory, cybernetics, game theory, decision theory, communication theory, and information theory[1] would seem to be highly appropriate for this purpose Such methods were all devised for dealing with

complex "goal-directed" systems, many of whose factors are not completely known and whose actions are governed in part by disturbances which are not specifically predictable.[2, 3]

The Basic Control System

If we take, for example, the simple directional orientation problem of a bird maintaining its flight course due north by means of a visual cue such as a north–south shoreline, the task can be readily conceptualized in terms of a control system. Whatever the system's flesh and blood components may be, it must be so constituted that it will minimize the difference between the bird's actual heading at any given moment and the reference direction, north (Fig. 57). The orienting mechanism will thus need to act as a regulator, adjusting the system's output so that a constant flight direction will be maintained despite buffeting by random gusts of wind or temporary diversions of the animal in pursuit of prey, in avoiding predators, and the like.

Note that the same animal orientation control system that performs as a regulator can act as a servomechanism if it responds to a variable reference signal rather than a steady one. Thus, in pursuing prey, the immediate orientation problem is to minimize the angle between the actively elusive prey, which provides the variable reference direction, and the pursuer's flight heading. This is, in fact, a servocontrol task.

To propose even the simplest model of our orienting control system in the bird, we must introduce one or more "black boxes" into the circuit. The black box is a fundamental concept in systems analysis and related disciplines.[1, 4] It comprises all or part of the system being analyzed and is considered as a single element for reasons of strategy, convenience, or experimental necessity. The black box has an input and an output in addition to its unspecified contents.

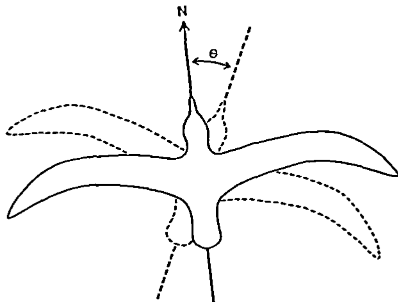


FIG. 57 Orienting bird flying north The broken lines indicate its relative position and flight direction when it is off course nearly 25° to the right heading NNE The angle θ thus represents the error between the intended course and the actual course at a given moment. An efficient mechanism controlling azimuth orientation would eliminate or minimize this error

the understanding of whose functional organization is really the ultimate scientific objective In approaching this goal, the systems analyst must first obtain the transfer function of the black box under study.

This transfer function is a mathematical model of the quantitative relation between the input and the output of the black box Such functions may be either stochastic and take the form of a conditional probability matrix, or deterministic and take the form of some differential, exponential, or other kind of equation. Usually, many transfer functions can be made to model a given input–output relation Selection of any particular one will depend, among other things, on the demonstrable relevance of the coefficients and vari-

ables in the model to the parameters of the biological system. Also the tractability of the equations is an important consideration. Thus, in the case of differential equations, these should, of course, be soluble, which generally requires that they be linear and have constant coefficients. Fortunately, many natural systems can be reasonably well modeled by equations having these restrictions.

For the biologist using systems analysis of this kind, the next step is to induce the nature of the black box's contents, employing the evidence of the transfer function to do so. The resulting white box,[1] which has its hardware and circuitry specified, must then be tested by further comparisons of its behavior with that of the corresponding black box. In this manner, the essential circularity of science is established by moving from data (black box behavior) to model (white box behavior), back to more extensive data, thence to improved models, and so on.[1, 5]

A Steering Control Model

While the northward orienting bird could be represented as a single black box with sensory inputs and motor outputs, a somewhat more specific model is desirable for even the simplest physiological consideration. Its essential elements can be diagrammed in a two-part circuit containing three black boxes, two in one part of the system and one in the other (Fig. 58).

The two major parts are a controlling system and a controlled system. The latter comprises complex neuromuscular mechanical and aerodynamic elements in the wings and tail, which steer the animal's flight; this part may be treated as a single black box with a complex transfer function, $G(s)$. The controlling system is considered to consist of two black boxes in series, one a strategy unit which acts as a sort of

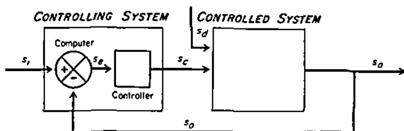


FIG 58 Basic control mechanism of the sort required by an orienting animal This minimal kind of system has two inputs, an out-

that constant adjustments are required to minimize output errors The controlled system in the bird orientation case would be the muscular, mechanical, and aerodynamic elements effecting steering action of the wings and tail The controller would be the central nervous system motor center that "commands" the effectors involved while the computer, acting as an error detector, would be some part of the information-processing system that activates that motor center in a fashion quantitatively related to the difference between the two inputs to the computer element s_i, input signal (if this is constant for appreciable periods, the system is a regulator, if it is variable, the system is a servomechanism), s_e, error signal, s_c, control signal, s_d, disturbances, s_o, output signal which is negatively fed back to the computer unit Note that the disturbances may influence many parts of the whole mechanism, but for simplicity they are shown here as just affecting the controlled system

computer, the other a controller which generates nerve impulses in the motoneurons innervating the controlled system. Each of these black boxes will, of course, have a transfer function of its own

A basic requirement for a control system is a means of comparing the actual output with the desired or reference output. This is usually provided by a closed-loop negative feedback which, in the present case, permits the strategy element in the controlling system to subtract the bird's actual

directional heading from the desired one northward. If there is a difference, this will give rise to an error signal, s_e, which is a function of the azimuth error angle θ and which is transmitted to the controller. The latter converts the sign and amplitude of the error signal into appropriate neuromotor steering signals, s_c, which act to eliminate the deviation from the intended course.

In the case of the bird orienting in azimuth, the behaviorally most interesting output of the neuromuscular system will be a steering force or turning tendency around the dorsoventral axis. The effect of this force is measured in terms of the actual heading of the animal, and the latter provides the signal s_o used for the negative feedback in our elementary diagram (Fig. 58).

One more important input element must be included in our generalized control circuit in order to represent random or other unpredictable disturbances acting on the system. A gust of wind, obstacle avoidance, pursuit of prey, and fatigue are examples of the kinds of contributing factor here. Such disturbances, whether of internal or external origin, may be most simply represented by an input s_d to the controlled system; both s_d and s_c will, of course, contribute to the system's output.

The inevitable presence of s_d in any practical control system composed of a number of parts and exposed to a complex environment is, in fact, the usual condition which requires the characteristic negative feedback loop of cybernetics. If noise of this sort were not present, a system once correctly oriented could be started on its way and continue indefinitely without directional errors. However, such errorless systems responses are practically nonexistent, and closed negative feedback loops are characteristic of control mechanisms in organisms as well as in machines. In addition, there are commonly present positive feedback loops (so organized that they do not lead to instability) and feed-ahead loops.

The Nature of the Reference Value

Note that in describing this elementary control system an objective, goal, or reference value has had to be introduced. Two comments on this are apropos. First, the reference values to which a control system regulates or follows arise outside the mechanism itself and in a sense constitute a weakness of systems analysis which has to take the goal as given. Second, when one speaks of such reference levels or objectives in an animal control system, no implication of human conscious purposive or teleological behavior should be read into it

The regulating system of the organism at the level here discussed no more has a conscious purpose than has a thermostat that maintains a room's temperature constant at some preset level Nevertheless for the living system, as for the man-made regulator, reference values are required, and they must be provided somehow In organisms they come from information storage elements which hold learned experience for subsequent reference, or from inherited experience in the form of genetic information, or both

This whole fund of information is selected in nature by the relative adaptiveness of the various available alternatives whenever the organism has one or more choices to make. Adaptiveness itself may be defined in terms of the survival value of any particular course of action [6, 7] Maladaptive choices will decrease the likelihood of the system's survival In critical cases, this in turn will lead to the death of the individual organism or to the extinction of populations, species, higher taxa, or even conceivably of all life, depending on the actual importance of the wrong choice Specifically, then, the adequacy of any animal orientation controlling system will be determined by its survival value Hence, the reference levels or objectives of the orienting mechanism must reflect this basic generality

The Relevance of Information Theory

In proposing a circuit for a steering control system (Fig. 58), the terminology and concepts of elementary cybernetics or control theory have been used to describe the way this model works. Furthermore, some ideas derived from information theory and communication theory have been introduced. Consequently, it is desirable to examine the relevance of these latter types of analysis in somewhat more detail.

To begin with, one almost automatically tends to speak nowadays of "signals" being transmitted over various pathways in all parts of such biological control circuits from input to output. The appropriateness of such usage, however, is more than colloquial, because the signals in question are closely analogous to the information that Shannon was dealing with in his development of the formalisms of communication theory. Consequently, deliberate application of the concepts and quantitative methods of this analytical technique should be fruitful in understanding an animal's steering control system.[8]

In fact, we shall see below that Shannon's definition of information, the concepts of channel capacity, optimal coding, decoding, signal-to-noise ratio, ambiguity, redundancy, and so on are all useful in analyzing and modeling the behavioral mechanism of orientation. Thus, if one considers such biological devices from an information theory point of view, two important basic principles emerge.

Quastler's Principle of Signatures

The first of these two has been called the Principle of Signatures by Quastler,[1, 9] who pointed out that it applies at many biological levels, from molecules to whole organisms and ecosystems. The principle states that, of the total information available to an organism in a given situation, only a

small fraction is actually used in making decisions or producing responses This key part of the total information Quastler called the signature.

At the molecular level, this principle may be illustrated by the biological properties of proteins. For instance, consider a large protein with a molecular weight of 100,000 or more, made up of several hundred amino acids whose kind, number, and complex arrangement must be highly specific. Yet, for this molecule's immunogenic properties only a small active site comprising a few amino acids is specifically critical, while the large amount of additional information in the protein has no direct role in this function.

Furthermore, if this same protein molecule were a hydrolytic enzyme, its enzymatic properties would undoubtedly reside in a different small site also comprising just a few amino acid groups. Thus, in Quastler's terms, these two active sites are the signatures for their respective functions

At the organizational level relevant to directional orientation of animals, this principle may be dramatically illustrated as follows The total sensory input rate in man, through his eyes, ears, taste buds, and all other exteroceptive sensory channels, has been estimated[10] to be about 10^7 bits/sec, where one bit is a binary logarithmic unit of information as defined by Shannon However, the rate at which decisions can be made to effect voluntary neuromuscular responses appears, on a similar basis, to be only 5–25 bits/sec, far lower than the estimated sensory inflow.

Such calculations show that only about one out of a million bits of information input can actually be decisive in the response output. Hence the Principle of Signatures must be operative here through a very restrictive bottleneck somewhere in the system's channel capacity More specifically, the processing of visual information in the orientation control system must strongly reduce the redundancy (or reduplication) of sensory data received before the bottleneck is reached

This must be done in an adaptive way so that no information is lost which is essential to making the right decisions for survival.

The Problem of Reliability

The second general principle that an information theory approach brings to the fore relates to the problem of producing a highly reliable overall system out of component parts which may be structurally or functionally unreliable. The task basically is one of reducing or eliminating errors in performance. At the molecular level this problem is exemplified by the still unknown mechanism whereby protein synthesis by organisms proceeds, as it does, with incredibly low error rates.[12] At the sense organ level, Helmholtz[13] recognized it nearly a hundred years ago when he enumerated the many physical optical shortcomings of the human eye. The cornea has scratches, the lens is not precisely shaped and may have cloudy blotches in it, the retina is inside-out, there are blood vessels between the stimulus source and the light-receiving elements, and so on. Yet Helmholtz realized, as we do, that the human eye is a much more reliable, highly effective total sense organ than the enumeration of such obvious defects in its various individual components would lead one to expect.

In the case of control circuits, a general way of producing reliable behavior out of unreliable parts is to duplicate components or systems in whole or in part. Such redundancy should function so that (1) a second system will take over in the event of failure of one, (2) the total information is divided into a series of partially overlapping parallel channels each of which mainly responds to one parameter of the total stimulus pattern, or (3) there are duplicate channels which ought to be carrying the same information. In the last case, absence of response in one unit at a time when a number of

others of the same sort respond can be safely interpreted on a probabilistic basis as a failure of that particular unit

Cowan[14, 15] has in fact demonstrated that the kinds of quasineuronal networks required to produce reliable automata made up of less reliable parts need both complexity and redundancy Hence, numbers of parallel channels are required with extensive interconnections both for excitation and inhibition As a result, there would be marked overlap in the input parameters carried by these parallel channels, and any given function would be transmitted to a number of places.

On the basis of these two principles we would anticipate that the processing of visual information relevant to an animal's orientation would show evidence both for marked reduction in redundancy to prevent overloading the channel capacity and for increased redundancy to provide the reliability in total performance required for survival As in the solution of many engineering design problems, the optimal biological system must provide the best compromise between these two basically antagonistic needs

Current Research Program

With the above discussion as the basis for a point of view, our current research program on visual responses in aquatic animals may now be considered Since interest has centered around orientation to polarized light, our work has dealt with animals which respond strongly to linear polarization Some research has been done with insects and cephalopods, but crustaceans have been used primarily Three quite different approaches to the problem have been followed. These should ultimately form a coherent body of knowledge which would permit us to understand the functional organization of oriented behavior. However, the correlations that can be made at the present stage of our work are to a considerable

extent still incomplete or hypothetical and many basic aspects of the system are not adequately known.

Our three approaches to the problem have been: (1) electrophysiological study of visual information processing in large decapod crustaceans, (2) study of phototactic responses to linearly polarized light, and (3) fine structure analysis of the receptor and nervous elements involved in the sensing and processing of visual information in an effort to understand the relation between input and output (1 and 2 above). These topics will be reviewed in order.

Compound eye physiology and visual signal processing have been studied most intensively in the swimming crab *Podophthalmus*, common in Hawaii (Fig. 59). Professor C. A. G. Wiersma of the California Institute of Technology and Dr. Brian M. H. Bush of the University of Cambridge, England, collaborated with the author on this research.[11, 16, 17] *Podophthalmus* was chosen as the experimental animal because its extraordinarily long eyestalks greatly facilitate single-fiber analysis of optic nerve signals. The great length of the proximal eyestalk segment (3–5 cm) provided an elongate optic nerve which can, in fact, be teased out into small bundles or single axons for electrical recording.

In this whole nerve there were undoubtedly many thousand afferent and efferent interneurons in addition to the motoneurons innervating the eyestalk muscles. In a crayfish having about 3,200 facets in its eye, a total of 19,800 axons was counted in the optic nerve.[18] Facet numbers have not been accurately counted in *Podophthalmus*, but we have estimated their number to be about 10,000. All the afferent fibers were parts of secondary or higher order units, since every primary visual fiber was believed to terminate in the most distal optic ganglion (Fig. 60). Since there were three additional more proximal interconnected optic ganglia, there was ample histological evidence that extensive processing of the primary visual responses could take place before the

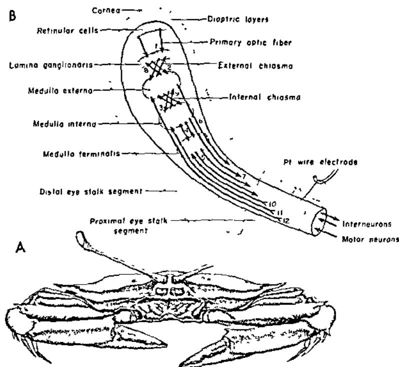


FIG 59 A Frontal view of the crab *Podophthalmus vigil* with its remarkably elongate eyestalks raised in a commonly observed position

B Diagrammatic longitudinal section of the distal part of the eyestalk showing the most likely neuronal connections between the retina, where the primary photochemical events occur, and the optic nerve, where recordings were made Four ganglia and two chiasmata occur in this pathway in which efferent interneurons have been hypothesized (as shown) to reach at least to the lamina ganglionaris but probably no farther (From Waterman et al [11])

information reached the level of our electrode Our recordings proved that this in fact was the case

In the first series of experiments, afferent responses to visual stimulation of the ipsilateral eye were recorded for single axons in the outer end of the proximal eyestalk segment Great care was necessary in making such preparations

stopped, visual responses in the optic nerve (propagated spikes) failed within a few minutes.

Responses of single afferent interneurons to a variety of visual stimuli showed that individual fibers differed markedly in their sensitivity to various parameters of the stimulus. In addition, some axons were spontaneously active (i.e responded steadily in the absence of obvious sensory stimulation), and others were mixed modality fibers responding to mechanoreceptor stimulation as well as to light.

e

Light enters at upper right through the corneal facets, passing downward and left to the retina deeper in the eye In this view the retina appears to consist of columnar clusters of retinular cells of which there are seven in each ommatidium The rhabdom, which is the most likely site of the visual pigment, lies centrally in the optic axis of each retinula The primary optic fibers are proximal axons of the retinular cells and pass through the basilar membrane, thence they run in bundles past well-developed blood vessels to the lamina

se in this

legree of

beneath

the retina; *bv*, blood vessels, which are plentiful and of several kinds, being particularly noteworthy just proximal to the basilar membrane and just distal to the lamina ganglionaris (*lg*), *c*, cornea, *cc*, crystalline cone, *cs*, crystalline cone stalk, *lg*, lamina ganglionaris, *pf*, primary optic fibers which have penetrated the basilar membrane and then run in bundles to *lg*, *rc*, retinular cells around rhabdom, *rh*, rhabdom axially located (Section prepared by Mabelita Campbell)

Lower Cross section through the retinal region of the compound eye of the mangrove crab *Goniopsis cruentata* roughly perpendicular to the optic axes of the retinulas concerned The retinulas are extraordinarily widely spaced from one another (as are those of *Podophthalmus*, above) and appear as flower-like cell clusters in which the petals are the seven retinular cells and the center is the axial rhabdom *rc*, retinular cells, *rh*, rhabdom, located in the retinula's optical axis and made up of layers of microtubules readily observable in electron micrographs (Figs 65-67) (Section prepared by Helen Hutchins)

Visual Data Processing

Of those neurons which gave simple responses and reacted only to light, there were three major categories. (1) Units signaling the duration and intensity of the stimulus in a manner closely similar to primary neurosensory cells (although latent periods could be quite long, e.g. 150 msec for some of the crab units). (2) On-off units which responded mainly to changes in light intensity, some of these specifically only to decreases or increases, others to both (comparable types of fiber from other decapods are illustrated in Fig. 61). (3) Movement-sensitive axons which responded trivially

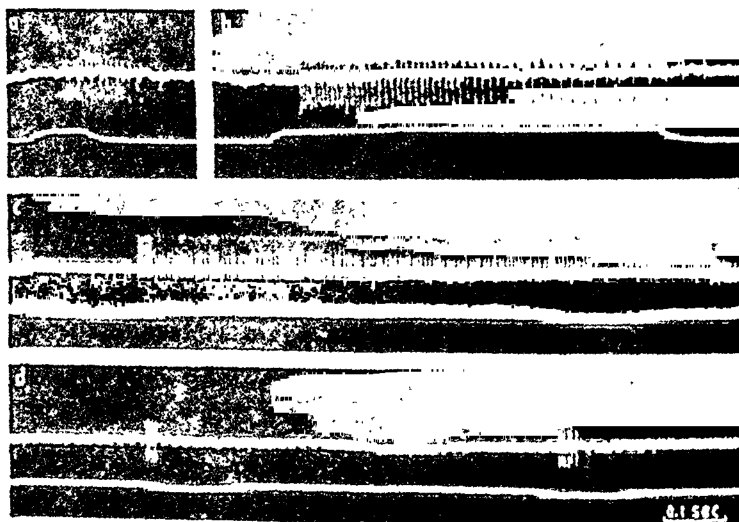


FIG. 61. Spike responses of several types of optic nerve afferent fibers. In the first three cases, records of single sustaining fibers in the spiny lobster *Panulirus argus* are shown. (a) Response to a 0.09-sec flash. (b) Response to a 0.59-sec flash. (c) Responses to shadows (lasting about 0.15 sec) from a swinging pendulum. (d) Responses of a single on-off fiber in the crab *Grapsus grapsus* to the same pendulum shadow stimulus used for c. In each recording, deflections of the lower trace indicate the flashes or shadows. (From Waterman and Wiersma.[179])

or irregularly to intensity levels or intensity changes and required a moving visual stimulus in their field for regular and high-frequency responses.

These movement units were more numerous than the other two types together and were differentiated among themselves in a number of ways. Some showed rapid adaptation so that the same movement repeated soon failed to stimulate or stimulated much less than the first or first few presentations (Fig. 62). The most rapidly adapting units of this kind were essentially novelty fibers, responding only to new movements in their field and requiring a minute or more to regain sensitivity between repetitions. Others adapted more moderately, and some only slightly, to repeated presentation. Different movement fibers were also specific for velocity (Fig. 63) or direction of target movement Consequently, there was a whole spectrum of specialized units of many kinds.

Such fiber inventories prove that extensive processing of visual information had indeed taken place between the primary event of light absorption and consequent impulse conduction in the optic nerve. This processing involved the distribution of the information into parallel overlapping neural channels differentially sensitive to particular stimulus parameters. All the fibers studied had large visual fields (30–180° or more), and no evidence was found for axonal tracts conducting highly localized positional information Similarly, no fibers responding strongly and specifically to polarized light were found. At present, we cannot say whether these two apparent deficiencies in optic nerve information were due to failure to record from the appropriate fiber categories or to inability to recognize the way in which these kinds of information were coded at the level studied.

A mathematical model of some interest for such a biological visual system may be seen in the methods of matrix algebra used to analyze the multivariate transformations of polarized

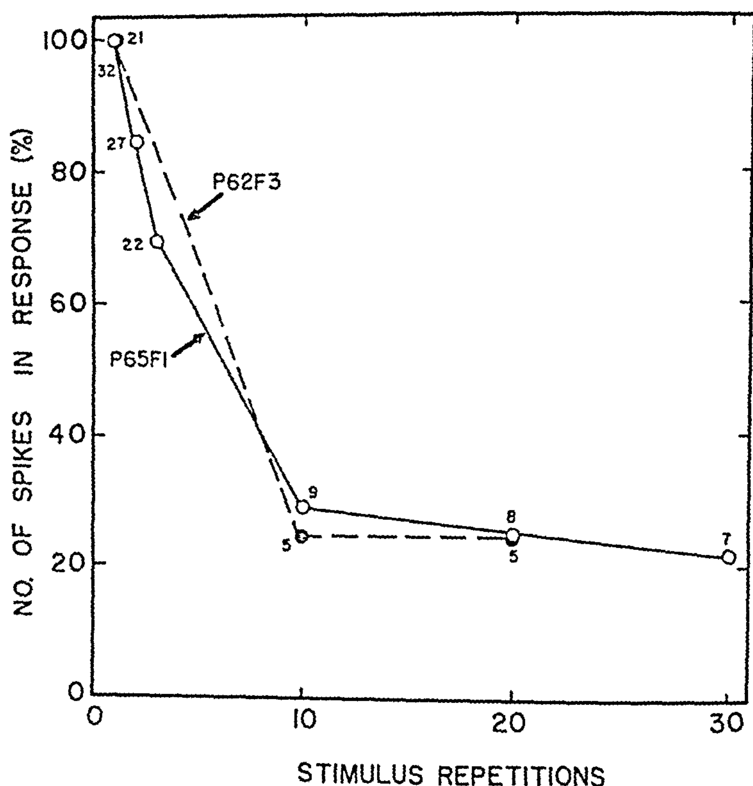


FIG. 62. Adaptation in the responses of two optic nerve fibers in the crab *Podophthalmus* stimulated by the repeated passage of a vertical stripe through their visual fields. Rapid and extensive initial decreases in the numbers of spikes per stimulus (figures next to the points) are evident, followed by sustained low-level responses. (From Waterman et al.[19])

light waves passing through complex optical systems.[20] To illustrate this point, one may paraphrase the general approach employed in the field of polarization optics in Mueller's and Jones' calculi as follows. The several relevant properties of the system's input are taken to be a column space vector $[V_i]$ comprising the corresponding number of stimulus parameters. These might be, in our *Podophthalmus*

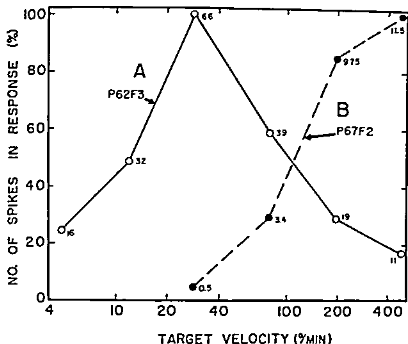


FIG. 63 Influence of target velocity on the movement responses of two optic nerve fibers in the crab *Podophthalmus* Fiber A's peak response occurred at an angular velocity barely detectable for Fiber B, while Fiber A's train of spikes had declined sharply at the velocities in excess of 400°/min necessary to get a high-level response from Fiber B. (From Waterman et al [111])

case, intensity, intensity modulation, polarization angle, directional characteristic, slow or fast movement, and so on

The interaction of this multidimensional stimulus input with functional properties of the primary sensory events could be modeled by an appropriate operation between this mathematical vector and a matrix representing the essential properties of the first processing element in the system, e.g the primary photoreceptor cells Subsequent processing would be modeled by operations with successive matrices $[M_i]$ of appropriate number and kind. Ultimately the result

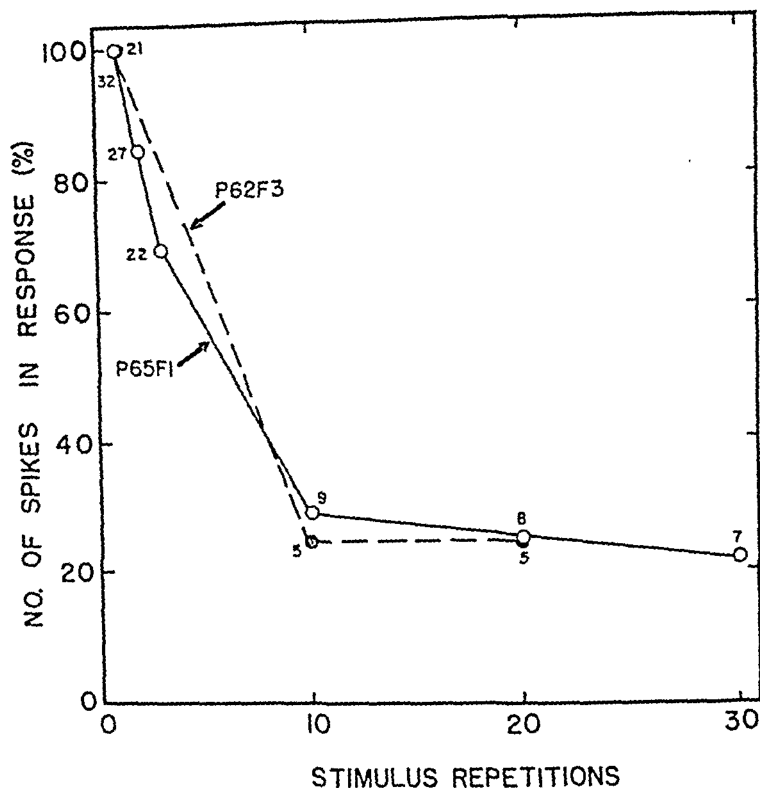


FIG. 62. Adaptation in the responses of two optic nerve fibers in the crab *Podophthalmus* stimulated by the repeated passage of a vertical stripe through their visual fields. Rapid and extensive initial decreases in the numbers of spikes per stimulus (figures next to the points) are evident, followed by sustained low-level responses. (From Waterman et al.[11])

light waves passing through complex optical systems.[20] To illustrate this point, one may paraphrase the general approach employed in the field of polarization optics in Mueller's and Jones' calculi as follows. The several relevant properties of the system's input are taken to be a column space vector $[V_s]$ comprising the corresponding number of stimulus parameters. These might be, in our *Podophthalmus*

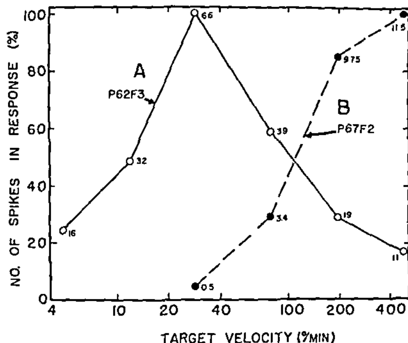


Fig 63. Influence of target velocity on the movement responses of two optic nerve fibers in the crab *Podophthalmus* Fiber A's peak response occurred at an angular velocity barely detectable for Fiber B, while Fiber A's train of spikes had declined sharply at the velocities in excess of 400°/min necessary to get a high-level response from Fiber B (From Waterman et al [111])

case, intensity, intensity modulation, polarization angle, directional characteristic, slow or fast movement, and so on

The interaction of this multidimensional stimulus input with functional properties of the primary sensory events could be modeled by an appropriate operation between this mathematical vector and a matrix representing the essential properties of the first processing element in the system, e.g the primary photoreceptor cells Subsequent processing would be modeled by operations with successive matrices $[M_i]$ of appropriate number and kind Ultimately the result

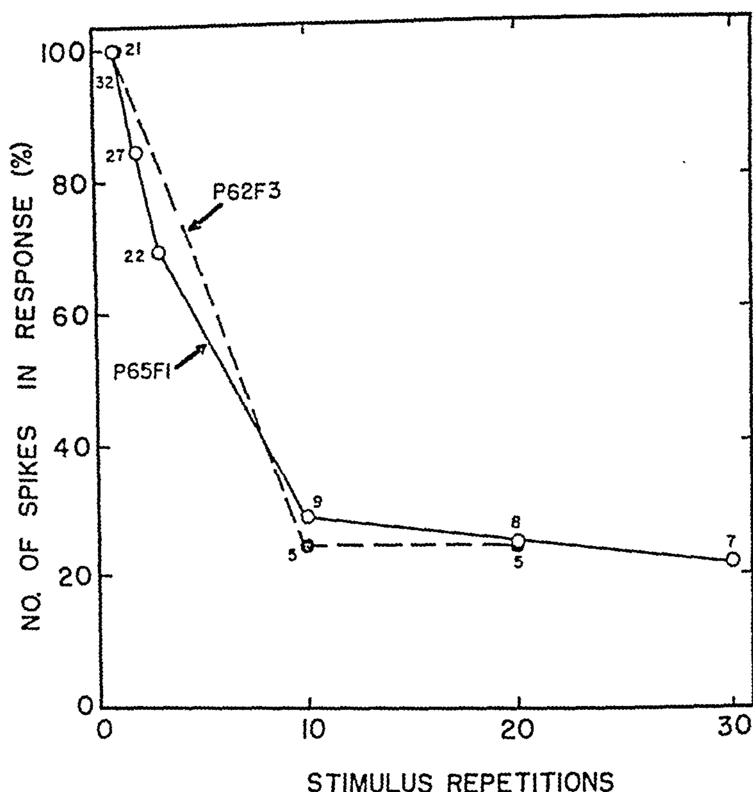


Fig. 62. Adaptation in the responses of two optic nerve fibers in the crab *Podophthalmus* stimulated by the repeated passage of a vertical stripe through their visual fields. Rapid and extensive initial decreases in the numbers of spikes per stimulus (figures next to the points) are evident, followed by sustained low-level responses. (From Waterman et al.[11])

light waves passing through complex optical systems.[20] To illustrate this point, one may paraphrase the general approach employed in the field of polarization optics in Mueller's and Jones' calculi as follows. The several relevant properties of the system's input are taken to be a column space vector $[V_x]$ comprising the corresponding number of stimulus parameters. These might be, in our *Podophthalmus*

Efferent Optic Nerve Traffic

In further series of our experiments with *Podophthalmus* the ipsilateral eye was surgically removed and efferent responses were measured in single units of the optic nerve on that same side. Previous work had suggested a considerable amount of efferent interneuron traffic in the eyestalks of other decapods[19] This suggestion was strongly supported by the later work which also showed, at least qualitatively, that all the different afferent visual fiber types obtained with ipsilateral illumination were matched by comparable efferent interneurons responding to contralateral light stimulation[16]

In addition, bending of claw or leg joints, touching of leg or body surfaces, and stimulation of the statocysts located in the antennule bases all produced corresponding efferent signals in the eyestalk nerve.[17] These results on *Podophthalmus* can be colloquially summarized by saying that each eyestalk was telling the other what was happening to it while the appendages and body were similarly communicating with both eyestalks [1]

This widespread exchange of sensory data seems quite different from a simple telephone switchboard model for nervous system function where a single primary projection area is usually considered *the* destination for specific afferent information. From the points of view of control theory and information theory, the observed situation is also of considerable interest because the multiple widespread transmission of sensory data could be involved in complex control networks, in information storage, or in redundancy to increase performance reliability. Much further work is required to test these possibilities and more specifically to explore in de-

1 Evidence that in the spiny lobster, *Panulirus*, efferent chemoreceptor information was transmitted from the ipsilateral antennule out to the medulla terminalis in the eyestalk[22] adds another sensory modality to our generalization

tail the significance of the extensive sensory outflow toward the optic ganglia. Our present research in this area involves attempts to understand the function of these complex peripheral nervous centers, starting with afferent visual processing between the retina and the optic nerve.

Azimuth Orientation

For this discussion, attention will be limited to our experiments on basic orientation in a horizontal plane. These involve studies of oriented locomotion relative to a horizontal directional light source and orientation in azimuth in response to a vertical beam of linearly polarized light.[2] The crucial behavioral parameter in such orientation is the turning tendency about the dorsoventral axis which tends to steer the animal toward or away from the directional light stimulus or toward one or a few basic directions relative to the plane of polarization. Note that this torque is not strictly represented by either s_e or s_o in Figure 58 but is in fact an effective motor output somewhere between them.

While phototactic and polarotactic turning tendencies have not yet been measured electrophysiologically, they have been determined by ingenious behavioral experiments in a variety of animals. The classic work was that of Holst[24] on fishes, while Schöne and others have studied the phenomena in crustaceans, insects, and other fishes.[25-35]

Interestingly enough, the data for directional light orientation indicate that the transfer function for the steady-state case is a simple trigonometric function where the torque $s_t = \sin \theta$, where θ is the angular deviation of the animal's orientation from the reference direction and s_t the system's out-

2. Following Jaffe's[18] use of the term *polarotropism* for oriented growth responses of plants to polarized light, the directed locomotor response of organisms to polarized light will be referred to as *polarotaxis*. Such terminology parallels the classic usage of phototropic and phototactic.

put measured as turning tendency. This is of itself a matter of great interest both from sensory and neurological points of view because there seem to be no known relevant functions that have simple trigonometric relations to light direction or presynaptic input.

In fishes, there is good evidence that the sinusoidal optical turning tendency induced by oblique directional illumination originates in the central nervous system rather than in the eye.[34] Only in species *(Pterophyllum scalare, Gymnocorymbus ternetzi)* in which the histological pattern of the retina is uniform throughout does such a regular simple relation exist between the retinal point stimulated and the force of the turning tendency. In most kinds of fish the retina shows regional variation in the density and distribution of visual cells and, correspondingly, irregularities in the optical turning tendency shown by the dorsal light reflex. Even in *Gymnocorymbus* such histological and behavioral lack of uniformity appears under scotopic conditions.[31] These teleost data prove that the number of retinal elements stimulated by a given illuminated area as well as the retinal location of the latter determine the intensity of the reflex torque.

On the other hand, a sinusoidal transfer function is not surprising for orientation to gravity, where it has also been demonstrated. In that case, the effective sensory stimulus is the shearing force of the statolith against the mechanoreceptive epithelium. Since this force will vary directly as the sine of the error angle for orientation to gravity, a sinusoidal righting tendency would seem to have a direct peripheral origin.[21] Note that these results further imply a direct linear relationship between mechanical input and reflex behavioral output.

Presumably for polarized light, as for directional light, the turning tendency would also be a sinusoidal function of the error angle. However because of symmetry about the *e*-vector, the relation would be $s_r = \sin 2\theta$. Direct evidence for this w

obtained in *Daphnia* by measuring the interaction of orientation to gravity and orientation to a horizontal beam of linearly polarized light.[33]

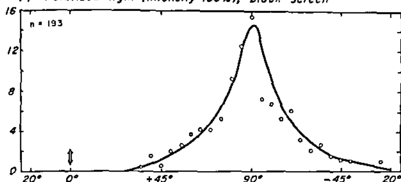
The actual patterns of polarotactic orientation observed in various animals and under different conditions in a vertical beam of linearly polarized light were of four different sorts (Table 6 in Ref. 36). The simplest responses involved in a half-circle, single-peaked orientation perpendicular to the *e*-vector (Fig. 64). Under other conditions or with different organisms, two-peaked patterns may be found either at 0° and 90° to the polarization plane or obliquely at +45° and −45°. Finally, all four peaks may be present in 180° (Fig. 64). Of course, these patterns differ not only in the number and direction of orientation peaks but also in the degree to which orientation occurs at all.

The Polarization Analyzer

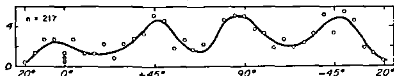
While much remains to be learned about polarotactic behavior itself and its relation to the normal field behavior of the animals concerned, the question of mechanism is particularly relevant to our present interest in systems analysis. To develop a working hypothesis for the observed polarized light orientation patterns, attention must be directed toward the nature of the receptor analyzer or analyzers involved. Since many of the basic facts remain to be discovered, several important assumptions must be made to devise a reasonable model.

Practically all data have pointed to a single retinula as the functional unit for specific polarized light analysis.[36] Despite several attempts at demonstration, no evidence has been found indicating that the corneal lens, crystalline cone, or cone stalk was the required analyzer.[37] The rhabdom is the next more central ommatidial element in the optic axis. This structure is directly in the path of the incoming light and has

DAPHNIA PULEX







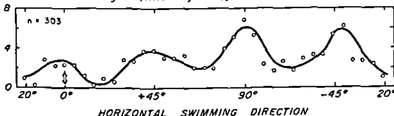


FIG 64 Oriented responses of *Daphnia pulex* in a vertical beam of linearly polarized light. With high-intensity unidirectional illumination single peaked orientation was marked *(A)* With reduced intensity *(B)*, decreased directionality of the overall illumination pattern *(C)*, or both *(D)*, four orientation peaks are in evidence (From Jander and Waterman [36])

been quite generally believed (although direct evidence is lacking) to be the site of rhodopsin. This is the photosensitive pigment type of decapod crustacean eyes as it is in insects, cephalopods, and in the rods of the great majority of vertebrates.[38]

If we wish to consider that the rhabdom may be both the photoreceptor organelle and the polarization analyzer, some knowledge of its fine structure is essential to support such hypotheses. In crustaceans, as in all animals known to be highly sensitive to polarized light, this retinal component is made up of thousands of regularly arranged microvilli or microtubules (Fig. 65). These are axial processes of the cluster of seven or eight neurosensory cells comprising each retinula in decapods. Normally, these tubules lie perpendicular to the retinula's optic axis and are quite uniform in size and arrangement for a given species.

Their diameters range in these Crustacea[39] from about 500 to 1,000 Å, as in insects[40, 41] with the order of magnitude about the same as the microtubules of *Octopus* (600 Å[42]), and spider rhabdoms (200–1,000 Å[43]). The lamellar disks of the outer segments of vertebrate visual cells are 100–250 Å thick, but the period of the stacked units is 100 Å or so greater.[44] Hence the dimensions of these complex membranous photoreceptor structures in vertebrates are similar to those of the finest invertebrate tubule systems reported.

In nearly all decapod crustaceans studied, as in a number of but not all insects,[41] and in those cephalopods that have been examined, all the microtubules in one retinula are aligned in two directions perpendicular to one another and to the optic axis. A special feature of the decapod and some other crustacean rhabdoms is that they are layered, a fact long known to the light microscopists.[45] Each of these layers has now been shown[39] to consist of closely parallel microtubules all contributed by three or four of the seven retinular cells and lying normal to those in adjacent layers (Fig. 65). Micro-



FIG. 65 Longitudinal section of the rhabdom of an ommatidium from the apposition-type compound eye of the giant land crab *Cardisoma guanhumi* The diameter of the rhabdom is indicated by R so that the optic axis of the ommatidium is nearly vertical with light entering from above and somewhat to the left The alternating layers of microtubules oriented at 90° in successive bands is quite clear Microtubules originate from the medial surface of the retinular cells (as can be seen in the case of many tubules parallel to the section) and extend about halfway across the rhabdom to meet their counterparts from cells on the opposite side B indicates the resulting boundary, C, bridge of retinular cell cytoplasm, V, large intracellular vacuoles around rhabdom (Electron micrograph by E Eguchi)

villi from a given retinular cell extend only halfway across their layer where they meet corresponding elements from a cell opposite; the axial ends of all tubules appear to be closed off like fingers of a glove (Figs. 66, 67). While the significance of the special layered arrangement is as yet unknown, the presence of two perpendicular sets of microtubules is quite likely a critical detail for polarotaxis.

To demonstrate this, another major assumption must be made, but this is one that seems reasonable by analogy with vertebrate rod and cone outer segments. There, the known receptor organelle consists of stacks of several hundred lamellar disks perpendicular to the optic axis. The photosensitive rhodopsin molecules are apparently arranged in a highly ordered quasicrystalline fashion relative to this multifold lamellar membrane system so that their longitudinal axes and hence their chromophores are parallel to the lamellae.

Evidence for this comes from the demonstrated strong dichroism of rod outer segments when examined in a direction normal to the optical axis[46-49]. This dichroism has been presumed to be of molecular origin since asymmetries of the rhodopsin molecule make it dichroic, i.e. absorbing light-wave trains with electric vectors vibrating parallel to its longitudinal axis more frequently than those oscillating transversely to it. Analysis of polarotropic responses of germinating plant spores has demonstrated that the relevant photoreceptor molecules are dichroic and specifically oriented relative to the cell wall.[50]

Note that the resulting overall dichroism of the rod outer segment is such that light passing through the eye in the usual direction for vision is not affected. This is quite consistent with the absence of strong polarized light sensitivity in the vertebrates. The weak polarized light sensitivity of the human eye manifested in Haidinger's brushes is most likely due to dichroism in the macular pigment which is present in the outer retinal layers of the fovea but not in the receptor



FIG. 66 Cross section of a *Cardisoma* rhabdom cut roughly perpendicular to the ommatidium's optic axis. Most of the area B cuts through only one of the banded layers of microtubules but, because the cut is somewhat oblique, the sharply perpendicular position of the microtubules of a second, neighboring layer is clear in area A as well as a probable third bank oriented like layer B. The retinular cells contributing to the rhabdom are numbered 1–7 with 1, 4, and
5 forming alternate microt
whereas cells 2, 3, 6, and 7
to those of the first group
arating retinular cells are at *i*, one of the large intracellular vacuoles at V, and pigment granules at P. (Electron micrograph by E. Eguchi.)

cells.[20] But the multiple tubular structure of the receptor organelle in arthropods and cephalopods would provide an analyzer effective for light incident in the usual direction for vision if the rhodopsin molecules were arranged in a quasicrystalline pattern as in the vertebrates.[23, 37, 51] Thus, if the rhodopsin molecules were oriented so that their chromophores were accurately parallel to the lipid–protein tubule membrane, overall light absorption should increase by a factor of 100 per cent.[51] The direct evidence for such a situation is not conclusive, but some support is available from

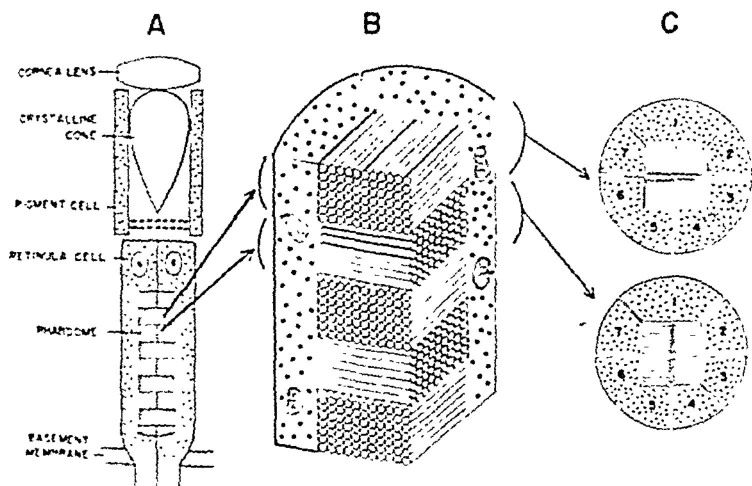


Fig. 67 Diagrams showing the structure of the rhabdom in the crayfish (*Procambarus clarkii*). Even though this is a well-developed superposition eye in Exner's classification, its rhabdom fine structure closely resembles that of the apposition eyes of the crabs studied. In A the ommatidium is shown shortened by omission of the elongate cone stalk, actually relatively longer here than in *Podophthalmus* (Fig. 60 top), as suggested by double broken lines; B indicates the microtubule packing pattern in a stereodiagram; C shows how the seven retinular cells of a retinula contribute to the rhabdom. Note the strong double peaking in microtubule orientation with maxima 90° apart. (From Eguchi.[29])

the reported dichroism of dipteran retinulas[52] and of cephalopod microtubule systems[53]

On the basis of this kind of hypothesized mechanism, the two perpendicular sets of decapod microtubules would give rise to light absorption proportional to $\sin^2\theta$ and $\cos^2\theta$ respectively, where θ is the angular difference between the e-vector of the incoming light and the direction of the more absorbing axis of the rhodopsin in the tubules Thus, the direction of the electric vibration of linearly polarized light is coded in the output of a single retinula as two trigonometrically related axon spike frequencies This kind of splitting the data analytically into sine and cosine components is a basic feature of Mittelstaedt's bicomponent theory of orientation.[54] Also, it is a mathematical feature of a well-known method of reducing statistical distributions on a circle to their mean vector (references in Waterman[55] and Markl[56]).[3]

Note that all such two-channel systems are different from the four-channel one originally hypothesized to account for polarized light sensitivity in the honeybee's eye[57, 58] Jander[33] has developed a functional scheme for an ommatidial analyzer mechanism based on the Frisch model In this scheme, two pairs of perpendicular analyzers were hypothesized to lie at 45° to each other A single output for each pair was suggested to follow the difference function $\cos 2\theta$ of their orthogonal inputs Discrepancies between the model and the structure of the rhabdom[59] were thought possibly to depend on specific regional differentiation of the retina such as that reported for *Drosophila*[60] However, the widespread occurrence of two-channel retinular analyzers as

3 Since this manuscript was completed, we have obtained electroretinographic evidence that two mutually perpendicular analyzer channels are, in fact, present in the retina of the crab, *Cardisoma* (Waterman and Horch, in preparation) These experimental data, obtained by selective adaptation, support the inferences made here on the basis of rhabdom fine structure (Figs 65 and 66)

well as the basic need to discriminate intensity and polarization,[36, 61] would seem to make this less likely. In fact, even in *Apis*, only two perpendicular sets of microtubules are reported to be present despite their location in four closely approximated rhabdomere pairs originating from eight retinular cells.[59] The repeated occurrence, in the polarized light-sensitive arthropods and cephalopods, of only two polarization analyzers in a single rhabdom raises some interesting problems relating to unambiguous *e*-vector discrimination and to polarotactic mechanisms.[41, 51]

Model for e-Vector Orientation

To devise a reasonable model for polarized light orientation, we may return to the experimentally measured turning

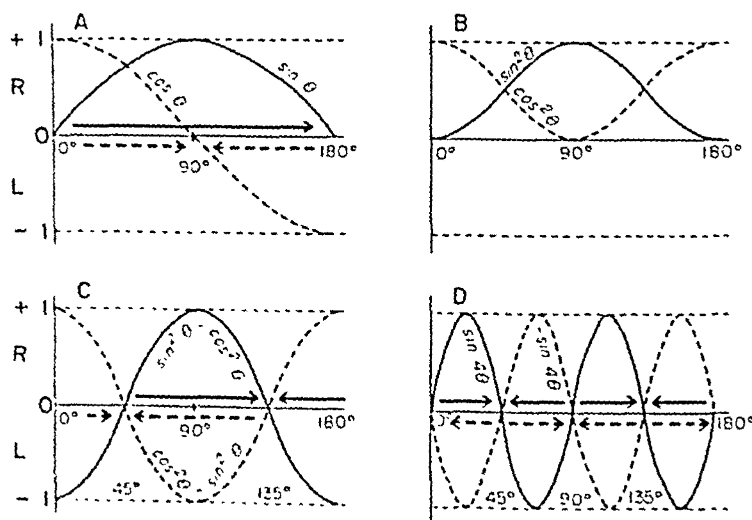


FIG. 68. A. Graphical model for a system showing sinusoidal turning tendencies with a period of 360°. The positive quadrants of the function are assumed to determine a right-turning tendency and the negative quadrants a turn to the left. Consequently, the arrows show the direction of the torques, and the amplitude of the

tendencies which control basic orientation to directional light and to linearly polarized light. In both cases, these are sinusoidal functions with the reference azimuths located at inflection points. For directional light the period is 360°, and the curve crosses the x-axis at 0 and 180° rising to +1 at 90° and falling to −1 at 270° (Fig 68).

Hence the animal's turning tendency would be nil at azimuths of 0° and 180° relative to the stimulating light, and these directions of locomotion would represent steady-state headings from which the animal would not actively turn. If we assume that, in the positive quadrants of the function, the animal turns right and in the negative quadrants left (of course the choice of which is which here is arbitrary for the model), the organism would turn right between 0° and 180° and left between 180° and 360° Thus, the steady state for

---

trigonometric function shows their intensity for any given azimuth direction In the case of sin θ, between 0° and 180° the organism [illegible]

azimuth of 180°

Thus a sin θ function would describe a negative phototaxis whereas −sin θ would describe a positively phototactic case Similarly, a system following a cos θ function would orient perpendicularly to the stimulus, as suggested by the broken arrows The well-known dorsal light reflex of swimming and flying animals is an example of such transverse orientation

B Sinusoidal functions representing the amount of polarized light absorbed by two dichroic elements oriented perpendicular to each other Note that the period is 180° and no negative values appear

C If the two functions in B are subtracted from each other, two curves may be obtained as shown here The turning tendencies shown by the arrows would produce steering at ±45° oblique to the e-vector

D If the outputs of the two systems are multiplied, one could obtain functions of 4θ which would give orientation at 0° and 90° as well as ±45°, depending on signs Such models should suggest experiments capable of testing the information-processing hypotheses involved

orientation at 0° (=360°) would be unstable since any inadvertent small deviation from that direction would lead to active turning still farther away from it. In contrast, 180° would be a stable orientation direction since accidental displacements would result in active turning back to 180°.

Thus the sin θ turning function describes the behavior of a negatively phototactic animal walking, swimming, or otherwise heading away from the directional light, provided the latter is at 0°. If the sign of the turning function were reversed ($-\sin \theta$), the model specified would apply instead to a positively phototactic animal with a stable orientation at 0°, heading directly toward the stimulus.

Comparable treatment of the polarized light case where the turning tendency varies as sin 2θ, leads to orientation perpendicular to the *e*-vector. Such a turning function could thus account for the single-peak orientation patterns commonly observed in *Daphnia* and many other forms (Jander and Waterman,[36] Table 6). Note that neither the ambiguity noted by Goldsmith[41] nor the restrictions suggested by Moody and Parriss[51] are relevant here because both positive and negative quadrants of the function are used for information and because the exact location of the *e*-vector is not "read out" instantaneously except when the animal is actually oriented in the reference direction.

As a result, such a mechanism would respond immediately, as do the organisms, by turning toward a specific basic heading relative to the *e*-vector through the smaller angle of the two available to it in 180°. The response of an animal which happened to be at the unstable inflection point when the polarization began would not be instantaneous, since a slight random deviation would be required before active turning could start. The significance of these factors in discrimination will become clear from the following.

The above model makes use of an overall transfer function relating stimulus *e*-vector direction and measured behavioral

turning tendency; it makes no use of the evidence for a two-channel analyzer and its optical properties. To do this we recall that the amount of light energy absorbed in the two channels will vary as $\sin^2\theta$ and $\cos^2\theta$ respectively.[4]

If these primary response intensities were added in subsequent neural processing, the result would be invariant with *e*-vector rotation but would change with overall light intensity. Hence a secondary neuron summing the two channels by additive convergence would provide information on light intensity independent of linear polarization, an important property required of the biological system.[26]

However, if the information in the two channels from a single retinula were subtracted by some neural process, the result would be a function of plus or minus $\cos 2\theta$ (depending on the order of subtraction). These curves if presented as turning functions could form the basis of orientation at $\pm 45°$ to the *e*-vector. Such a mechanism has already been suggested[33] for the hypothesized four-channel model of Frisch and Autrum.

A simple but not unreasonable means for accomplishing the kinds of turning tendencies required in the present case can be devised by using the negative as well as the positive phases of the difference curve (Fig. 69). This model comprises excitatory and inhibitory innervation controlling motor systems evoking right and left turning tendencies. In this way, the information transmitted by the two optical channels, which for a fixed *e*-vector input are in themselves ambiguous for the quadrants 0–90° and 90–180°, is processed in a simple manner which could provide unambiguous steering control through a semicircle.

---

4 Note that the logarithmic transformation typically found between stimulus intensity and receptor potential has not been used in this model because the experimentally observed turning tendencies appear to be direct and not transformed sinusoidal functions. Furthermore overall linear transformations have been reported in an insect visual system (Reichardt[62] p. 367) and in other sensory systems.[24, 63–65]

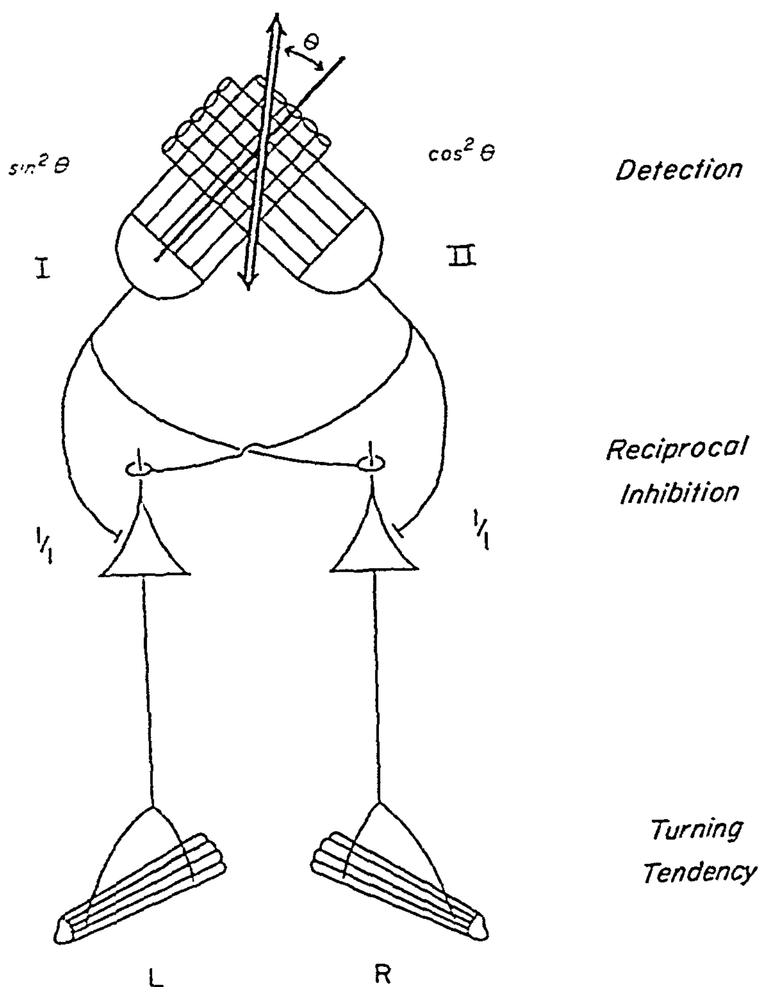


FIG. 69. Model of a two-channel polarized light detecting system which could produce right and left turning tendencies like some of those experimentally observed. Because the microtubule patterns have alternating bands arranged with their long axes perpendicular to those of the neighboring layers, the outputs would vary as $\cos^2 \theta$ for Channel I and $\sin^2 \theta$ for Channel II as indicated. Retinular cells 1, 4, and 5 would follow one function while 2, 3, 6, and 7 would follow the other. If reciprocal excitation and inhibition occurred as suggested in the figure, turning tendencies like those obtained in experiments could be the model's outcome. The looped synapses are assumed to be inhibitory, the others excitatory with one-to-one transmission ratios.

Two peaked orientation curves could be obtained from such a model if the outputs of the two channels were multiplied.[5] In this case an output function in cos 4θ is obtained. This could be processed to yield orientation peaks at 0° and 90° or +45° and −45° where the functions cross the abscissa. Hypotheses about a sin 8θ function which would directly provide the commonly observed four-peaked response would seem too speculative at this point to be fruitful. For instance, we do not yet know experimentally whether all four directional alternatives are simultaneously available to the orienting animal or whether switching occurs from one to another at intervals Fresh study of such details would quite likely illuminate the question of mechanisms involved. Consequently, further elaboration of the model should await new experimental data

Future Work on Orientation

In concluding this review two points should be made. One is that we recognize that the present state of this work, both experimentally and analytically, is still a humble one Much remains to be done before the overall problem can reach a reasonably sophisticated level It is clear, for example, that little of the dynamic data which a control system engineer would require to establish a transfer function has yet been obtained for our biological orienting system

Thus, effective analysis of a regulator or servomechanism would require quantitative knowledge of its responses to both steady-state and transient inputs The former most simply should be sinusoidal inputs of all relevant frequencies, while corresponding response measurements should comprise output gain and phase data as a function of such frequencies. A step function or rapid pulse would be the appropriate input for determining transient characteristics

5 Experimental evidence for multiplicative interaction of neural channels in an insect visual system has been analyzed by Reichardt[62] (p. 359).

Furthermore, adequate analysis of control system behavior demands that input–output relations be measured both in the intact system with the essential negative feedback loop closed and in the system modified so that this loop has been opened. Relatively few reflex or other behavioral systems have been analyzed in these ways,[66, 67] and our polarized light orienting mechanism not only requires more carefully controlled quantitative inputs of appropriate kinds but also more definitive identification of system components, including the feedback loop. Interesting examples of applications of such methods to other biological problems have been given, among others, by Hassenstein,[68] Mittelstaedt,[54] and Grodins.[3]

More Complex Orientation

In delimiting the current scope of this work it is also desirable to repeat that only basic light reactions have been considered. These comprise either positive or negative phototaxis and the four alternative directional reactions to a beam of linearly polarized light. Such relatively simple and stereotyped outputs will ordinarily provide only the underlying reference directions for more complex kinds of orientation. Almost universally, animals can use such visual reference directions for steering in any azimuth.

In such cases, the orientation is a light compass reaction or menotaxis for which the basic sinusoidal turning tendency must be modified to shift the direction of stable orientation as required. The small amount of relevant evidence available as proposed in Mittelstaedt's hypothesis[54] is not yet widely conclusive on whether the phase of the basic transfer function is somehow altered or whether a biasing signal shifts the position of the sinusoid along the y-axis so the mean of the function is no longer zero.[29, 33, 69, 70]

Furthermore, the menotactic response, from annelids on

up through the animal kingdom to mollusks, arthropods, and vertebrates, commonly uses celestial bodies (usually the sun but *including the moon and perhaps the stars*) to provide phototactic reference directions Such orientation, which may be called astrotaxis, most likely is superimposed on the underlying basitactic and menotactic mechanisms, it requires two further complex components in the control system.

One of these components is knowledge of the path through the sky of the celestial body concerned; such knowledge may be learned from experience by the individual or transmitted to it genetically The second component required is a clock mechanism which will permit effective correlation of (1) the observed sky position of the reference celestial body and (2) *azimuth direction on the earth's surface*. Analysis of such mechanisms is beyond the scope of this paper, but they must be mentioned to put basic light orientation in a larger perspective. The same should be said for navigation of the sort shown, for example, in some insects, fishes, and birds which somehow can determine their position on the earth's surface and then, starting from this, set a long-distance homing course.

Multidisciplinary Approach

The *second major point to be made in conclusion concerns* the author's convictions that even relatively simple biological problems often require a high-powered multidisciplinary approach. Our example of visual orientation demonstrates not only that a number of disciplines within biology (e.g electrophysiology and electron microscopy) are necessary even to begin the search for adequate solutions but also that a number of academically remote methods of study such as control theory, computer techniques, information theory, and mathematics of various kinds, including multivariate statistics, are required

Thus, the biologist who initiates research on living systems, in addition to being trained in the life sciences, must also know at least enough about other probably relevant areas of science to understand the basic requirements for an adequate explanation of the phenomena under study. Undoubtedly the most important prerequisite here is an ingrained familiarity with the fundamentals of mathematics. These come as near as anything to a *lingua franca* for rigorous and highly conceptual thinking about natural phenomena.

The need for such a basic language would seem to be at least as great for biologists as for any scientists, particularly in view of the complex multivariable nature of living systems. Furthermore, some formalism like systems theory is essential for dealing with such complex systems, which involve many still unknown components and interactions. If, in reviewing our current research on visual orientation with the aid of rather elementary systems analysis, we have been able to induce or to add some conviction to the above generalities, the effort involved will surely have been more than worthwhile.

Addendum

Because of the considerable progress made since this review was originally delivered as a lecture in the fall of 1964 and submitted for publication a year later, an addendum summarizing the main trends of recent work seems appropriate. Substantial gains have been made not only in the general area of systems analysis in biology but also in the more specific domain of our own research: visual orientation and the relevant information processing in the arthropod retina. Both areas in fact have been vigorous contributors to the publication explosion.

While a detailed résumé of the whole field would be out of place here, some examples may be cited of books whose publication is symptomatic of the recent marked acceleration

in the growth of theoretical and mathematical biology. These may be roughly categorized under: biomathematics,[1a–3a] computer applications,[4a–6a] cybernetics and control systems,[7a–13a] bioengineering,[14a–16a] and systems analysis.[17a] In the last field a Symposium on Systems Analysis in Biology was held in Cleveland in October 1966; its proceedings are in press.[18a] The author's most recent discussion of this topic is his contribution to that volume.[19a]

In the field of visual information processing, particularly in the compound eye, much work has been reported. Three symposia in 1965 were either partly[20a, 21a] or wholly[22a] devoted to this topic and several important independent papers have been published. Among the latter, direct demonstration has now been made that the visual pigment in the rhabdomeres of dipteran insect eyes is actually dichroic in situ[23a] Significant correlations, again in flies, have been made between the functional pattern of receptor elements and ommatidial optics[24a, 25a] as well as between retinular cell patterns and the complex projection of their first-order neurons onto the second-order cells in the outer ganglion, the lamina ganglionaris[26a–28a] Some progress along these lines has been made in other insects[29a–31a] and in crustaceans[32a–34a] Considerable new work has also been done on the aberrant compound eye of *Limulus* (review by Wolbarsht and Yeandle[35a]).

Our own research has continued to be concentrated on crustaceans and as in the past has included fine structural[36a, 37a] electrophysiological (referred to in footnote on p 353 but now published as Waterman and Horch[38a]) and behavioral (Daumer and Waterman, unpublished; summary in Waterman[37a]) approaches to the problem The major advance has been to prove by more than one technique that the two-channel dichroic analyzer model (Fig 69) proposed above is indeed valid for polarized light perception by decapod crustaceans.

Two methods of measuring selective adaptation in crab compound eyes have been utilized. In the first of these the effect of a steady polarized adapting light was tested by measuring extracellularly electroretinograms (ERG's) in response to brief test flashes polarized successively in various directions.[38a] These experiments on the land crab *Cardisoma* proved that selective adaptation by steady polarized light did readily occur but was maximal in two orthogonal directions and minimal for adaptation at intermediate directions 45° oblique to the most effective ones.

This proved that there were indeed two perpendicular polarized light analyzer channels as hypothesized. Further study of rhabdom fine structure showed that these channels were oriented parallel to the two sets of microvilli in the rhabdom. Consequently the two directions of maximal polarization sensitivity corresponded with those predicted by the model. However, since the ERG's were recorded with extracellular electrodes, some ambiguities remained concerning the cellular basis of the analyzer channels.

Such uncertainty has now been resolved by experiments which took advantage of the fact that long-lasting light adaptation produces marked changes in the numbers of several classes of cytoplasmic elements in the photoreceptor cells of the spider crab retinula.[39a] It has been quantitatively demonstrated that six hours of light adaptation to polarized light evoked cytoplasmic adaptation in specific cells of the retinula in characteristic patterns indicating two perpendicular channels.[100] Thus the three retinular cells (Nos. 1, 4, and 5) with horizontal microvilli were selectively more strongly light adapted by a horizontal *e*-vector, while the remaining four retinular cells (Nos. 2, 3, 6, and 7) with vertical microvilli were more affected by a vertical *e*-vector.

These new data, therefore, are also completely compatible with the two-channel analyzer model proposed above. In this they confirm by a quite different means the conclusions

drawn from the electrophysiological experiments.[380] In addition, however, they establish the cellular basis for the mechanism in a manner not directly possible previously. Thus the data prove that both channels are present in each retinula and comprise certain specific cells therein. Also they prove that the major axis of polarized light absorption is parallel to the long axis of the microvilli of the cells concerned. Again this new fact is in accord with the two-channel hypothesis and as a result supports the assumption that the dichroism of oriented visual pigment molecules in the rhabdom is the basis for *e*-vector perception. Molecular orientation predominantly parallel to the microvillus' long axis would be expected from the data on vertebrates and had therefore been predicated in the model

REFERENCES

1 H QUASTLER, General Principles of Systems Analysis in *Theoretical and Mathematical Biology*, T H Waterman and H J Morowitz, Eds, Blaisdell, New York (1965) p 313

2 J D TRIMMER, *Response of Physical Systems*, Wiley, New York (1950)

3 F S GRODINS, *Control Theory and Biological Systems* Columbia Univ Press, New York (1963)

4 N WIENER, *Cybernetics, or, Control and Communication in the Animal and the Machine*, 2d ed, M I T Press Wiley, New York (1953)

5 H MARGENAU, *Open Vistas*, Yale Univ Press, New Haven (1961)

6 T. H. WATERMAN, Comparative Physiology, in *The Physiology of Crustacea*, T H. Waterman, Ed Academic Press, New York (1961) vol 2, p 521.

7. R LEVINS, Genetic Consequences of Natural Selection, in *Theoretical and Mathematical Biology*, T. H Waterman and H J Morowitz, Eds, Blaisdell, New York (1965) p.371

8 C E SHANNON and W WEAVER, *The Mathematical Theory of Communication*, Univ of Illinois Press, Urbana (1963)

9 H. QUASTLER, *The Emergence of Biological Organization*, Yale Univ Press, New Haven (1964)

10. H. QUASTLER, Studies of Human Channel Capacity, in *Information Theory*, London Symposium on Information Theory. 1955, C. Cherry, Ed., Academic Press, New York (1956) p.361.
11. T. H. WATERMAN, C. A. G. WIERSMA, and B. M. H. BUSH, Afferent Visual Responses in the Optic Nerve of the Crab, *Podophthalmus, J. Cell. Comp. Physiol., 63*, 135 (1964).
12. C. B. ANFINSEN, *The Molecular Basis of Evolution*, Wiley, New York (1959).
13. H. VON HELMHOLTZ, *Handbuch der physiologischen Optik*, 1st ed., Part 3, Voss, Leipzig (1866).
14. J. D. COWAN, The Engineering Approach to the Problem of Biological Integration, in *Nerve, Brain and Memory Models*, N. Wiener and J. P. Schadé, Eds., Elsevier, Amsterdam (1963) p. 22.
15. J. D. COWAN, Redundant Automata as Models of Neuron Assemblies, in *Information Processing in the Nervous System*, R. W. Gerard and J. W. Duyff, Eds., *Proc. Intl. Union Physiol. Sci.* (22d Intl. Congr., Leiden, 1962), *3*, 397 (1964).
16. C. A. G. WIERSMA, B. M. H. BUSH, and T. H. WATERMAN, Efferent Visual Responses of Contralateral Origin in the Optic Nerve of the Crab, *Podophthalmus, J. Cell. Comp. Physiol., 64*, 309 (1964).
17. B. M. H. BUSH, C. A. G. WIERSMA, and T. H. WATERMAN, Efferent Mechanoreceptive Responses in the Optic Nerve of the Crab, *Podophthalmus, J. Cell. Comp. Physiol., 64*, 327 (1964).
18. R. F. NUNNEMACHER, G. CAMOUGIS, and J. H. MCALEAR, The Fine Structure of the Crayfish Nervous System, *5th Intl. Congr. for Electron Microscopy, 1962*, S. S. Breese, Jr., Ed., Academic Press, New York (1962) p.N-11.
19. T. H. WATERMAN and C. A. G. WIERSMA, Electrical Respons[illegible] in Decapod Crustacean Visual Systems, *J. Cell. Comp. Physiol* *61*, 1 (1963).
20. W. A. SHURCLIFF, *Polarized Light; Production and Use*, Ha[illegible] Univ. Press, Cambridge (1962).
21. J. L. STEPHENSON and A. P. JONES, Linear Methods in [illegible] in *The Cullowhee Conference on Training in Biomathem[illegible]* Lucas, Ed., Institute of Statistics, North Carolina S[illegible] Raleigh (1962), p.200.
22. D. M. MAYNARD and H. DINGLE, An Effect of Ey[illegible] on Antennular Function in the Spiny Lobster, [illegible] *Z. vergl. Physiol., 46*, 515 (1963).
23. L. F. JAFFE, Tropistic Responses of Zygotes of [illegible] Polarized Light. *Exp. Cell Res., 15*, 282 (1958).

24. E. von Holst, Die Arbeitsweise des Statolithenapparates bei Fischen. *Z. vergl. Physiol.*, *32*, 60 (1950).

25. H. Schöne, Die Lichtorientierung der Larven von *Acilius sulcatus L.* und *Dytiscus marginalis L.*, *Z. vergl. Physiol.*, *33*, 6 (1951).

26. H. Schöne, Zur optischen Lageorientierung ("Lichtrücken-orientierung") von Dekapoden. *Naturwiss.*, *23*, 552 (1952).

27. H. Schöne, Über den optischen Lageapparat der Krebse, i *Verhandl. deutsch. zool. Ges.*, *19*, 52 (1955).

28. H. Schöne, Die Lageorientierung mit Statolithenorganen und Augen. *Ergeb. Biol.*, *21*, 11 (1959).

29. H. Schöne, Optisch gesteuerte Lageänderungen (Versuche an Dytiscidenlarven zur Vertikalorientierung), *Z. vergl. Physiol.*, *45*, 590 (1962).

30. H. Schöne, Über die Arbeitsweise der Statolithenapparate bei Plattfischen. *Biol. Jahresheft*, *4*, 135 (1964)

31. W. Braemer, Verhaltensphysiologische Untersuchungen am optischen Apparat bei Fischen. *Z. vergl. Physiol.*, *39*, 374 (1957

32. R. Jander, Die optische Richtungsorientierung der rote Waldameise (*Formica rufa L.*). *Z. vergl. Physiol.*, *40*, 162 (1957

33. R. Jander, Grundleistungen der Licht- und Schwereorientierung von Insekten. *Z. vergl. Physiol.*, *47*, 381 (1963).

34. H. Bogenschütz, Vergleichende Untersuchungen über die optische Komponente der Gleichgewichtshaltung bei Fischen. *Z. vergl. Physiol.*, *44*, 626 (1961)

35. U. Bässler, Zum Einfluss von Schwerkraft und Licht auf die Ruhestellung der Stabheuschrecke (*Carausius morosus*). *Z. Naturforsch.*, *17b*, 477 (1962).

36. R. Jander and T. H. Waterman, Sensory Discrimination between Polarized Light and Light Intensity Patterns by Arthropods. *J. Cell. Comp. Physiol.*, *56*, 137 (1960).

37. K. Stockhammer, Die Orientierung nach der Schwingungsrichtung linearpolarisierten Lichtes und ihre sinnesphysiologischen Grundlagen. *Ergeb. Biol.*, *21*, 23 (1959)

38. H. J. A. Dartnall, The Identity and Distribution of Visual Pigments in the Animal Kingdom. in *The Eye*, H. Davson, Ed. Academic Press, New York (1962), vol. 2, p. 367.

39. E. Eguchi, "The Structure of the Rhabdom and Action Potentials of Single Retinula Cells in Crayfish," Ph.D. thesis, Kyushu Univ. (1964).

40. H. Fernández-Morán, The Fine Structure of Vertebrate and Invertebrate Photoreceptors as Revealed by Low-Temperature Electron Microscopy, in *The Structure of the Eye*, G. K. Smelser, Ed., Academic Press, New York (1961), p. 521.

41. T. H. Goldsmith, The Visual System of Insects, in *The Physiology of Insecta*, M. Rockstein, Ed., Academic Press, New York (1964), vol. 1, p. 397.

42. J. J. Wolken, Retinal Structure. Mollusc cephalopods: *Octopus, Sepia; J. Biophys. Biochem. Cytol.*, *4*, 835 (1958).

43. B. Baccetti and C. Bedini, Research on the Structure and Physiology of the Eyes of a Lycosid Spider. I. Microscopic and Ultramicroscopic Structure, *Arch. Ital. Biol.*, *102*, 97 (1964).

44. F. S. Sjöstrand, Fine Structure of Cytoplasm: The Organization of Membranous Layers, in *Biophysical Science—A Study Program*, J. L. Oncley, Ed., Wiley, New York (1959), p.301.

45. G. H. Parker, The Retina and Optic Ganglia in Decapods, Especially *Astacus; Mittheil. zool. Station Neapel, 12*, 1 (1895).

46. W. J. Schmidt, Doppelbrechung, Dichroismus und Feinbau des Aussengliedes der Sehzellen vom Frosch, *Z. Zellforsch. Mikroskop. Anat.*, 22, 485 (1935).

47. W. J. Schmidt, Polarisationsoptische Analyse der Verknüpfung von Protein- und Lipoidmolekeln, erläutert am Aussenglied der Sehzellen der Wirbeltiere, *Pubbl. Staz. zool. Napoli*, *23* (Suppl), 158 (1951).

48. E. J. Denton, The Contributions of the Orientated Photosensitive and Other Molecules to the Absorption of the Whole Retina, *Proc. Roy. Soc. London, 150B*, 78 (1959).

49. G. Wald, P. K. Brown, and I. R. Gibbons, Visual Excitation: A Chemo-anatomical Study, in *Biological Receptor Mechanisms*, Symp. Soc. Exp. Biol., Cambridge Univ. Press, Cambridge, *16*, 32 (1962).

50. L. F. Jaffe and H. Etzold, Orientation and Locus of Tropic Photoreceptor Molecules in Spores of *Botrytis* and *Osmunda*, *J. Cell Biol.*, *13*, 13 (1962).

51. M. F. Moody and J. R. Parriss, The Discrimination of Polarized Light by *Octopus:* A Behavioural and Morphological Study, *Z. vergl. Physiol.*, *44*, 268 (1961).

52. L. Giulio, Elektroretinographische Beweisführung dichroitischer Eigenschaften des Komplexauges bei Zweiflüglern, *Z. vergl. Physiol.*, *46*, 491 (1963).

53. M. F. Moody, Photoreceptor Organelles in Animals, *Biol. Rev. Cambridge Phil. Soc.*, *39*, 43 (1964).

54 H. MITTELSTAEDT, Control Systems of Orientation in Insects, *Ann Rev Entomol*, *7*, 177 (1962).

55. T H. WATERMAN, The Analysis of Spatial Orientation, *Ergeb Biol*, *26*, 98 (1963)

56. H MARKL, Geomenotaktische Fehlorientierung bei *Formica polyctena* Forster, *Z vergl Physiol*, *48*, 552 (1964)

57. H AUTRUM and H STUMPF, Das Bienenauge als Analysator fur polarisiertes Licht, *Z Naturforsch*, *5b*, 116 (1950)

58 K. VON FRISCH, Die Sonne als Kompass im Leben der Bienen, *Experientia*, *6*, 210 (1950)

59 T. H. GOLDSMITH, Fine Structure of the Retinulae in the Compound Eye of the Honey-Bee, *J Cell Biol*, *14*, 489 (1962)

60 R DANNEEL and B ZEUTZSCHEL, Ueber den Feinbau der Retinula bei *Drosophila melanogaster*, *Z Naturforsch*, *12b*, 580 (1957)

61 K VON FRISCH, M LINDAUER, and K DAUMER, Ueber die Wahrnehmung polarisierten Lichtes durch das Bienenauge, *Experientia*, *16*, 289 (1960)

62 W. REICHARDT, Nervous Processing of Sensory Information, in *Theoretical and Mathematical Biology*, T H Waterman and H J Morowitz, Eds , Blaisdell, New York (1965), p 344

63 S S STEVENS, The Psychophysics of Sensory Function, in *Sensory Communication*, Walter A Rosenblith, Ed , M I T Press, Wiley, New York (1961), p 1

64 G WENDLER, Laufen und Stehen der Stabheuschrecke *Carausius morosus* Sinnesborstenfelder in den Beingelenken als Glieder von Regelkreisen, *Z vergl Physiol*, *48*, 198 (1964)

65 V B MOUNTCASTLE, G F POGGIO, and G WERNER, The Neural Transformation of the Sensory Stimulus at the Cortical Input Level of the Somatic Afferent System, in *Information Processing in the Nervous System*, R. W Gerard and J W Duyff, Eds , *Proc Intl Union Physiol Sci* (22nd Intl Congr , Leiden, 1962), *3*, 196 (1964)

66 L STARK, Vision Servoanalysis of Pupil Reflex to Light, in *Medical Physics*, O Glasser, Ed , Year Book, Chicago (1960), vol 3, p 702

67 D VARJÚ, Ueber den Pupillenreflex auf mono-und binoculare dynamische Lichtreize (Abstr) *Pflugers Arch Ges Physiol*, *281*, 88 (1964)

68 B HASSENSTEIN, Die bisherige Rolle der Kybernetik in der biologischen Forschung, *Naturwiss Rundschau*, *13*, 3 (1960)

40. H. Fernández-Morán, The Fine Structure of Vertebrate and Invertebrate Photoreceptors as Revealed by Low-Temperature Electron Microscopy, in *The Structure of the Eye*, G. K. Smelser, Ed., Academic Press, New York (1961), p. 521.

41. T. H. Goldsmith, The Visual System of Insects, in *The Physiology of Insecta*, M. Rockstein, Ed., Academic Press, New York (1964), vol. I, p. 397.

42. J. J. Wolken, Retinal Structure. Mollusc cephalopods: *Octopus, Sepia; J. Biophys. Biochem. Cytol.*, *4*, 835 (1958).

43. B. Baccetti and C. Bedini, Research on the Structure and Physiology of the Eyes of a Lycosid Spider. I. Microscopic and Ultramicroscopic Structure, *Arch. Ital. Biol.*, *102*, 97 (1964).

44. F. S. Sjöstrand, Fine Structure of Cytoplasm: The Organization of Membranous Layers, in *Biophysical Science—A Study Program*, J. L. Oncley, Ed., Wiley, New York (1959), p.301.

45. G. H. Parker, The Retina and Optic Ganglia in Decapods, Especially *Astacus; Mittheil. zool. Station Neapel*, *12*, 1 (1895).

46. W. J. Schmidt, Doppelbrechung, Dichroismus und Feinbau des Aussengliedes der Sehzellen vom Frosch, *Z. Zellforsch. Mikroskop. Anat.*, 22, 485 (1935).

47. W. J. Schmidt, Polarisationsoptische Analyse der Verknüpfung von Protein- und Lipoidmolekeln, erläutert am Aussenglied der Sehzellen der Wirbeltiere, *Pubbl. Staz. zool. Napoli*, 23 (Suppl), 158 (1951).

48. E. J. Denton, The Contributions of the Orientated Photosensitive and Other Molecules to the Absorption of the Whole Retina, *Proc. Roy Soc. London*, *150B*, 78 (1959).

49. G. Wald, P. K. Brown, and I. R. Gibbons, Visual Excitation: A Chemo-anatomical Study, in *Biological Receptor Mechanisms*, Symp. Soc. Exp. Biol., Cambridge Univ. Press, Cambridge, *16*, 32 (1962).

50. L. F. Jaffe and H. Etzold, Orientation and Locus of Tropic Photoreceptor Molecules in Spores of *Botrytis* and *Osmunda*, *J. Cell Biol.*, *13*, 13 (1962).

51. M. F. Moody and J. R. Parriss, The Discrimination of Polarized Light by *Octopus*: A Behavioural and Morphological Study, *Z. vergl. Physiol.*, *44*, 268 (1961).

52. L. Giulio, Elektroretinographische Beweisführung dichroitischer Eigenschaften des Komplexauges bei Zweiflüglern, *Z. vergl. Physiol.*, *46*, 491 (1963).

53. M. F. Moody, Photoreceptor Organelles in Animals, *Biol. Rev. Cambridge Phil. Soc.*, *39*, 43 (1964).

54 H. MITTELSTAEDT, Control Systems of Orientation in Insects, *Ann Rev. Entomol*, *7*, 177 (1962)

55 T. H. WATERMAN, The Analysis of Spatial Orientation, *Ergeb Biol*, *26*, 98 (1963)

56. H MARKL, Geomenotaktische Fehlorientierung bei *Formica polyctena* Forster, *Z. vergl. Physiol*, *48*, 552 (1964).

57. H. AUTRUM and H. STUMPF, Das Bienenauge als Analysator fur polarisiertes Licht, *Z. Naturforsch*, *5b*, 116 (1950).

58. K. VON FRISCH, Die Sonne als Kompass im Leben der Bienen, *Experientia*, *6*, 210 (1950)

59 T. H. GOLDSMITH, Fine Structure of the Retinulae in the Compound Eye of the Honey-Bee, *J Cell Biol*, *14*, 489 (1962)

60. R DANNEFL and B. ZEUTZSCHEL, Ueber den Feinbau der Retinula bei *Drosophila melanogaster*, *Z. Naturforsch*, *12b*, 580 (1957)

61 K VON FRISCH, M LINDAUER, and K. DAUMER, Ueber die Wahrnehmung polarisierten Lichtes durch das Bienenauge, *Experientia*, *16*, 289 (1960)

62. W. REICHARDT, Nervous Processing of Sensory Information, in *Theoretical and Mathematical Biology*, T H Waterman and H. J. Morowitz, Eds, Blaisdell, New York (1965), p 344

63 S S STEVENS, The Psychophysics of Sensory Function, in *Sensory Communication*, Walter A Rosenblith, Ed, M I T Press, Wiley, New York (1961), p.1.

64. G WENDLER, Laufen und Stehen der Stabheuschrecke *Carausius morosus* Sinnesborstenfelder in den Beingelenken als Glieder von Regelkreisen, *Z vergl Physiol*, *48*, 198 (1964)

65. V. B MOUNTCASTLE, G F POGGIO, and G WERNER, The Neural Transformation of the Sensory Stimulus at the Cortical Input Level of the Somatic Afferent System, in *Information Processing in the Nervous System*, R W Gerard and J W Duyff Eds, *Proc. Intl Union Physiol Sci* (22nd Intl Congr., Leiden, 1962) *3*, 196 (1964).

66 L. STARK, Vision Servoanalysis of Pupil Reflex to Light, in *Medical Physics*, O Glasser, Ed, Year Book, Chicago (1960) vol. 3, p 702

67. D VARJÚ, Ueber den Pupillenreflex auf mono-und binoculare dynamische Lichtreize (Abstr) *Pflugers Arch Ges Physiol*, *281*, 88 (1964)

68. B HASSENSTEIN, Die bisherige Rolle der Kybernetik in der biologischen Forschung, *Naturwiss Rundschau*, *13*, 3 (1960)

69. R. JANDER, Menotaxis und Winkeltransponieren bei Köcherfliegen (Trichoptera), *Z. vergl. Physiol.*, *43*, 680 (1960).
70. R. JANDER, Insect Orientation, *Ann. Rev. Entomol.*, *8*, 95 (1963).

NEW ADDITIONAL REFERENCES

1a. S. R. SEARLE, *Matrix Algebra for the Biological Sciences*, Wiley, New York (1966).
2a. G. R. STIBITZ, *Mathematics in Medicine and the Life Sciences*, Yearbook, Chicago (1966).
3a. N. T. J. BAILEY, *The Mathematical Approach to Biology and Medicine*, Wiley, New York (1967).
4a. R. S. LEDLEY, *Use of Computers in Biology and Medicine*, McGraw-Hill, New York (1965).
5a. R. W. STACY and B. D. WAXMAN, *Computers in Biomedical Research*, Vol. I, Academic Press, New York (1965).
6a. H. ZIMMER, ed., *Computers in Psychophysiology*, Thomas, Springfield, Ill. (1967).
7a. M. MAXFIELD, A. CALLAHAN, and L. J. FOGEL, *Biophysics and Cybernetic Systems*, Spartan, Washington (1965).
8a. N. WIENER and J. P. SCHADÉ, eds., *Cybernetics of the Nervous System*, Elsevier, Amsterdam (1965).
9a. W. S. YAMAMOTO and J. R. BROBECK, *Physiological Controls and Regulations*, Saunders, Philadelphia (1965).
10a. L. R. YOUNG and L. STARK, *Biological Control Systems: A Critical Review and Evaluation*, NASA Rept. No. CR-190, Washington (1965).
11a. H. KALMUS, ed., *Regulation and Control in Living Systems*, Wiley, London (1966).
12a. H. T. MILHORN, JR., *The Application of Control Theory to Physiological Systems*, Saunders, Philadelphia (1966).
13a. J. H. MILSUM, *Biological Control Systems Analysis*, McGraw-Hill, New York (1966).
14a. F. ALT, ed., *Advances in Bioengineering and Instrumentation*, Vol. I, Plenum, New York (1966).
15a. K. ENSLEIN and J. F. KINSLOW, eds., *Data Acquisition and Processing in Biology*, Vol. 4, Pergamon, Oxford (1966).
16a. D. M. GREEN and J. A. SWETS, *Signal Detection and Psychophysics*, Wiley, New York (1966).

17a K. E. F. Watt, ed., *Systems Analysis in Ecology*, Academic Press, New York (1966).

18a M. D. Mesarovic, ed., *Symposium on Systems Approach in Biology*, Springer, Berlin (in press)

19a. T. H. Waterman, Systems Theory and Biology—View of a Biologist, in *Symposium on Systems Approach in Biology*, M. D. Mesarovic, ed., Springer, Berlin (in press)

20a L. Frisch, ed., Sensory Receptors, in *Cold Spring Harbor Symposium on Quantitative Biology, 30* (1966).

21a P. W. Nye, ed., *Information Processing in Sight Sensory Systems*, California Institute of Technology, Pasadena (1966)

22a C. G. Bernhard, ed., The Functional Organization of the Compound Eye, Wenner-Gren Center International Symposium Series, Vol. 7, Pergamon, Oxford (1966).

23a H. Langer, Nachweis dichroitischer Absorption des Sehfarbstoffes in den Rhabdomeren des Insektenauges, *Z. vergl. Physiol.*, *51*, 258 (1965)

24a D. M. Vowles, The Receptive Fields of Cells in the Retina of the Housefly *(Musca domestica)*, *Proc. Roy. Soc. Lond.*, *B*, *164*, 552 (1966)

25a K. Kirschfeld, Die Projektion der optischen Umwelt auf das Raster der Rhabdomere im Komplexauge von *Musca*, *Exp. Brain Res.*, *3*, 248 (1967)

26a O. Trujillo-Cenoz, Some Aspects of the Structural Organization of the Intermediate Retina of Dipterans, *J. Ultrastruct. Res.*, *13*, 1 (1965).

27a O. Trujillo-Cenóz, Some Aspects of the Structural Organization of the Arthropod Eye, in *Cold Spring Harbor Symposium on Quantitative Biology, 30*, L. Frisch, ed., 371 (1966)

28a V. Braitenberg, Patterns of Projection in the Visual System of the Fly. I. Retina-lamina Projections, *Exp. Brain Res.*, *3*, 271 (1967)

29a R. R. Bennett, J. Tunstall, and G. A. Horridge, Spectral Sensitivity of Single Retinula Cells of the Locust, *Z. vergl. Physiol.*, *55*, 195 (1967)

30a S. R. Shaw, Simultaneous Recordings from Two Cells in the Locust Retina, *Z. vergl. Physiol.*, *55*, 183 (1967)

31a J. Tunstall and G. A. Horridge, Electrophysiological Investigation of the Optics of the Locust Retina, *Z. vergl. Physiol.*, *55*, 167 (1967)

32a. J. Hámori and G. A. Horridge, The Lobster Optic Lamina I. General Organization, *J. Cell Sci.*, *1*, 249 (1966).

33a. J. Hámori and G. A. Horridge, The Lobster Optic Lamina II. Types of Synapse, *J. Cell Sci.*, *1*, 257 (1966).

34a. G. A. Horridge, Perception of Polarization Plane, Colour and Movement in Two Dimensions by the Crab, *Carcinus*, *Z. vergl. Physiol.*, *55*, 207 (1967).

35a. M. L. Wolbarsht and S. S. Yeandle, Visual Processes in the *Limulus* Eye, *Ann. Rev. Physiol.*, *29*, 513 (1967).

36a. E. Eguchi and T. H. Waterman, Fine Structure Patterns in Crustacean Rhabdoms, in *The Functional Organization of the Compound Eye*, C. G. Bernhard, ed., Wenner-Gren Center International Symposium Series, Vol. 7, Pergamon, Oxford (1966).

37a. T. H. Waterman, Polarotaxis and Primary Photoreceptor Events in Crustacea, in *The Functional Organization of the Compound Eye*, C. G. Bernhard, ed., Wenner-Gren Center International Symposium Series, Vol. 7, Pergamon, Oxford (1966).

38a. T. H. Waterman and K. W. Horch, Mechanism of Polarized Light Perception, *Science*, *154*, 467 (1966).

39a. E. Eguchi and T. H. Waterman, Changes in Retinal Fine Structure Induced in the Crab *Libinia* by Light and Dark Adaptation, *Z. Zellforsch.*, *79*, 209 (1967).

40a. E. Eguchi and T. H. Waterman, Cellular Basis for Polarized Light Perception in the Spider Crab, *Libinia*, in preparation (1967).

12. THE SELECTION OF SPACE EXPERIMENTS*

By W. H. Pickering
California Institute of Technology

As we move ahead in the space age, flights and projects are becoming increasingly complex. The variety of scientific experiments being flown is greatly increasing. I should like to discuss one phase of the program, the selection of space experiments.

When the older members among my readers were graduate students, the conduct of a scientific experiment seldom involved elaborate equipment or, if it did, the scientist expected to build the equipment with his own hands. A small machine shop, some optical instruments, a vacuum pump, a galvanometer, a spectroscope, and the physicist was ready to undertake his research. In other branches of science, the basic requirements were equally simple. A nostalgic remembrance of these good old days was a song that was popular with physicists at the end of World War II. To quote "Take away your billion dollars, every volt I make is pure."

Of course, not all science was done with "love and string and sealing wax" to quote another line from the same song. There were signs of an inevitable growth toward big science. Telescopes were getting larger. Scientific expeditions to the far corners of the globe were becoming popular. But World War II accelerated the process enormously because, for the

* The 1965 Procter Prize award address

first time, the federal government became a major patron of science.

The scientist has always needed a patron to supply him with funds and laboratories. A century ago, the patron was a rich man or a university. In the early years of this century, the foundations became the patrons of science, and now the federal government is playing the major role with industry, with the foundations supplying only about 25 per cent of the funds expended in the United States for scientific research.

Whether it be cause or effect, the result of the shift in patronage has been that the nature of scientific research has changed to utilize the larger monetary resources now available. Experiments have become increasingly complex, requiring ever more elaborate equipment, and teams of scientists and engineers to build the equipment and conduct the experiment. In some cases, the resources required have become so large that the government is the only possible patron.

In this evolutionary development of scientific patronage, we should remember that the scientist has always had the problem of convincing his patron that the research was worth supporting. Sometimes he has had to select a patron whose interests encompassed the research program. At other times he has had to choose a research to satisfy the desires of the patron. For example, Tycho Brahe cast horoscopes for his sovereign, the King of Denmark, and many a medieval scientist sought for the means of turning base metal into gold to make his patron rich.

With the federal government becoming a patron of science, some new factors are apparent, particularly regarding the support of very large projects. The expenditure of public funds is a matter of public record and concern. Therefore the Congress, and indeed the public at large, scrutinizes scientific proposals, and if resources are to be made available for a project, it will be with the express approval of the Con-

gress Furthermore, the funds required for some experiments are large enough to make a significant socioeconomic impact on a sector of the economy, and therefore various pressure groups may try to influence the decision.

Another factor is the experiment itself. Some scientific projects transcend the interests of a narrow group of scientific specialists and arouse the interest of lay people everywhere.

It is not surprising, therefore, that discussions of government support for a given project include considerations of "defense," "technological advancement," "cold war status," "industrial fallout," and similar catch phrases. Indeed, without an element of some of these factors, a large scientific research program cannot hope to have popular support Public understanding of all scientific research is at best rather vague, so that when the public discover that a large sum of their tax money is being used for a research project, they want to associate this work with some concrete, easily visualized objective rather than pure scientific research Unfortunately, there have been examples of much public misunderstanding because debates on the relative merits of certain projects, which have taken place quite properly within the scientific community, have been translated into the public domain. It is not always realized that discussions at the previous level of discourse may be used for quite different purposes at other levels. Perhaps we have to accept this as one of the consequences of greater public scrutiny, at least until there is a much broader understanding of science within the public and particularly within the Congress.

There is another side of governmental support of science which should be noted. The government supports a great deal of technological development as an essential activity of many departments, particularly of the Defense Department. This requires a broadly based advancement of the state of the art and a continuous search for new applications from

basic research. It follows, therefore, that support of scientific research is a natural outgrowth of this interest. In these areas, it is a case of the government patron seeking a scientist to do research with a predetermined objective. In other words, we still have the analog of the medieval patron seeking an alchemist to make him rich.

My purpose is to discuss some of the specific problems associated with a very large governmental scientific effort, namely, the space program. Let us remember that this program was established by Act of Congress in the immediate post-Sputnik period of 1958. The Act was the result of a public demand that the government respond to the Sputnik challenge. As finally passed, it set up a new agency, the NASA, with a charter which can be broadly stated as the exploration and exploitation of space. The budget for the program has now leveled off at about 5 billion dollars annually. The largest part of this money is the cost of the *Apollo* program, which was established by President Kennedy in May 1961.

Considered as a program for the scientific exploration of space near the earth, the NASA program illustrates very well the problems that both the sponsoring agency and the scientific community face with large, expensive, scientific projects that use government funds. The agency faces the problem of selecting a proper, well-balanced program that utilizes the available resources in a manner that forms a compromise between the interests of the various groups concerned in the results and which can also command the support of the Congress. The individual scientist who is selected to perform an experiment faces the problem of doing this experiment in a difficult physical environment and with the full scrutiny of the public as well as his scientific peers.

The NASA has made a firm decision to advance its program of exploration on a broad scientific front. This decision is supported by the space committees of the Congress. It has

resulted first in a series of scientific satellites designed to circle the earth in various orbits, and to examine the near-earth space environment and the appearance of the earth and the universe as seen from above the atmosphere. A second class of experiment is the deep-space or interplanetary *environmental monitoring program*. Finally, there is the detailed exploration of the moon and the planets. These three classes of scientific missions are being conducted within a total program which also includes the development of manned space flight, including the *Apollo* project, and the applications of satellites to meteorology and communications

The three classes of scientific missions have required the development of launching vehicles, spacecraft, and communications networks These have been designed to satisfy the requirements of most of the experimenters

Launching sites at both East Coast and West Coast test ranges have been developed so that satellite orbits of any inclination to the equator can be obtained Spacecraft that can be maintained in specified orientations in space are available, and we have the ability to land instruments gently on the moon Thus, NASA has provided the tools for a large variety of scientific experiments in space Perhaps not all the constraints have been removed, but the remaining ones tend to be those which not even NASA resources can eliminate For example, the data transmission rate from a spacecraft near or on the planet Mars will be severely limited for some time yet From the point of view of the scientific community, it should be quite apparent that NASA has undertaken and solved most of the engineering problems underlying the successful scientific exploration of space

In order to arrive at a specific set of flight missions, NASA has sought advice from the Space Science Board of the National Academy, and from several summer study groups The agency has also established a set of advisory committees to Dr Newell, Associate Administrator for Space Science

and Applications. It has then, with its field laboratories, selected a set of missions consistent with the financial and technical resources available. To obtain the necessary funds, the program, of course, has required the annual approval of the Congress.

Inevitably, the program which has evolved over the past seven years has not received the unanimous support of all scientists. Many of those not directly interested in space research have complained at the vast sums appropriated for NASA. Others within the program have complained at the distribution of funds among the various projects. However, viewing it dispassionately, we must agree that it has been a sound program, consistent with the intent of the Act and, also, that it has generated a great deal of new knowledge.

The selection of the actual experimenters for a mission is a step that is fraught with difficulties. NASA has made it a policy to announce its plans as widely as possible and to invite scientists who wish to be experimenters to make proposals. It has then gone through a selection process to arrive at the final groups of experimenters for any given mission. Without going into details regarding this selection, let me point out that the process does recognize that these space missions are truly national efforts, and therefore the experimenters are selected on a national basis rather than from one laboratory or university. The selection also recognizes that the conditions of a space flight impose rigid constraints on equipment, and therefore the state of development of potential flight hardware is very important.

From the point of view of a scientist desiring to carry out an actual flight experiment, there are several considerations he must not overlook. Let us review some of these matters.

First, the experiment is going to be very expensive. For example, the *Mariner* flight to Mars had a total program cost of about 100 million dollars. If this total cost were charged against data collected on the flight, it would come to about

4 dollars per second or 350 thousand dollars per day. Consequently, an experiment should not be carried on a flight unless it has a very high probability of working and of returning significant scientific data.

To make the point still stronger, it should be appreciated that many of these space experiments will have only a single opportunity to work Other experimenters may be selected for the next flight, or, if it is a planetary mission, the flight opportunities may be so infrequent as to invalidate the reason for an experiment

A second but related matter is the long lead time associated with space experiments. In some of the more complex missions, experiment selection may have to occur two or three years before flight Consequently, an experimenter must commit himself for a long period to a piece of research that might give him no data. A launching failure, a spacecraft failure, or even a failure of his own equipment could occur, and he would be left with nothing but a piece of prototype flight hardware

Sound engineering of rockets, spacecraft, and scientific experiments is the only insurance against these problems All the projects of "big science," whether in space or other fields, are dependent upon the quality of the engineering that goes into the equipment In fact, about 90 per cent of the budget of these projects typically goes to engineering and only 10 per cent to science The quality of the engineering, from system design to fabrication and test, is obviously of supreme importance This is particularly so in space projects because the complete system is operated only once, and that is the actual flight when everything must work.

Engineering control of a project requires that the scientific experimenter recognize two problems He must have a close rapport with the engineers so that the spacecraft will indeed fly in such a way as to give him the desired data, and he must understand and live within the constraints placed on his

equipment by the engineers. These range all the way from dollars and schedules to physical and electrical interfaces with the spacecraft and to test and documentation requirements. Sometimes they might appear onerous to a scientist used to performing isolated experiments within his own laboratory. However, they must be regarded as requirements arrived at as a result of long, hard experience and, if the mission is to succeed, they must be taken seriously.

A third matter which is new to many potential experimenters is that of public attention. Space science is conducted in the full spotlight of public scrutiny. The scientist is spending the public's money; he is conducting a spectacular experiment; he is using rockets and computers and worldwide tracking networks to collect his data; and he must expect the public to be interested. It is true that, as the years pass, and satellites are launched with great regularity and few spectacular accidents, the public is becoming blase, but the more complex missions still attract worldwide attention and the experimenters should anticipate the attentions of press and radio.

With all the engineering emphasis that must be present in a major space flight project, there is sometimes a concern expressed that the science has been submerged by the engineering glamor. Gordon MacDonald, in a recent speech before the American Physical Society, suggested that much of the current United States space effort is confusing glamorous technological achievements with fundamental science. I am sure that such confusion does not exist within responsible management at NASA headquarters. It probably does exist within a sector of the press and certainly within a sector of the public. In any case, it is up to the scientific community to see that there is in fact no such confusion. Developing the tools for large-scale space exploration has indeed put NASA into a position where the mere technological achievement has been so spectacular as to have public appeal regardless of the

total objective of the flight However, the increasing sophistication of press and public is taking care of the problem. A flight that fails to return some significant scientific data is quickly written off.

Now that the fundamental engineering difficulties have been resolved, NASA needs to devote the same kind of effort to science mission and system design that was previously expended in spacecraft system engineering. This will call for a close working relationship between the engineers who have developed the technology of system design and the scientists who have devised the experiments to be flown. Some of the more complex spacecraft, the *Mariner* for example, illustrate what can be done with a closely integrated design. Structure, power consumption, thermal balance, telemetry requirements, trajectory requirements, and spacecraft flight attitude were all elements in the experiment and instrument designs The result was a mission and a spacecraft which were uniquely matched to the particular experiments aboard Designs were frozen in all details ten months before flight. The problem of developing an appropriate set of mission objectives within the weight constraints determined by the launching rocket forced the designers to make this very finely tuned design. But the lesson is there for future missions which may not be so constrained A careful synthesis of the complete system, spacecraft plus science plus mission, will inevitably lead to a more efficient overall project with the maximum probability of success

For longer-range objectives, such as the exploration of Mars by a series of spacecraft flights, I conclude that systems engineering concepts must also be applied to the complete program The experiments, the spacecraft, the launching rockets, the communications networks must all be developed in a consistent manner if our resources are to be used most effectively This will require long-term commitments and stable plans Perhaps in an art that is developing so rapidly,

it is not possible to reach a stable optimum, but we should try, and we should plan for a flexibility that can absorb both technological and scientific surprises.

In conclusion, let me remind you that the scientific value of the national space effort is conditional upon the broadest possible base of participation by the scientific community. The space program is a national commitment and, by opening up an entirely new area to scientific investigation, it presents a unique opportunity to science. Already, very significant work has been accomplished, but the opportunities are growing every year as more sophisticated flights are being undertaken. As I have indicated, participation in the flight experiments may present some different problems to the scientist. However, the body of experience both within NASA and elsewhere is now such that he can face these problems with confidence. It should also be remembered that the flight experiments are only one part of the total space effort. Scientists are needed to conduct ground experiments which support and supplement the flight experiments. New ideas for improved missions, spacecraft, and instruments all need to be exploited.

The space program offers unlimited opportunities for imaginative scientific thinking. The tools are there, the resources are available, and the solar system is waiting to be explored.

INDEX

A. D. Little, Inc., 286
Academia and industry, 1-19, and establishment of RESA, 1-3, roles in catalysis research, 2-4, 12-16, and radio astronomy, 5-6, and nuclear science, 6-8, research on waves and particles, 8-9, tools and techniques, 9-12, and future scientific progress, 16-19, references, 19
Adler, Alfred, 320
Aiken, Howard, 12
Air Products and Chemicals Co., 312 313
Alabama Drydock & Shipbuilding Co., 288
Alabama Gas Co., 312
Alaska, 317
Algeria natural-gas liquefaction, 298, 299, 300, 301, 303, 304, proposed Mediterranean pipeline from, 305
Allfrey, V. G., 252, 253
Allison, A. C., 79, 80
American Education Association, 115
American Glass Research Co., Inc., 212
American Messer Corp., 313
American Physical Society, 380
Andrade, E. N. deC., 199
Andronikashvili, 23, 32, 38
Apollo spaceship program, 376, 377
Arzon, A. S., 199
Astronomy, radio, 5-6
Australia, 306
Autrum, H., 357
Averbach, B. L., 172
Bacterial cell, structure and possibility of artificial synthesis, 133-70, DNA, RNA, and their location, 137-43, inferences regarding cell-structure preliminaries to synthesis, 143-47, ordered synthesis, 147-56, schematic picture of, 156-63, ways of synthesizing, 163-67, references, 168-70
Badger, A. E., 209
Badger, G. M., 86, 87
Badische Anilin u. Soda Fabrik, 4
Baker, T. C., 193
Baker, W. O., 16
Bardeen, John 11
Bassham, J. A., 10
Battson, E. F., 285
Baudelaire, quoted 60
Bayou Long, La., 285, 286
Beare, Aleeza C., 130
Becker, R. S., 81
Becquerel, 7
Beebe, R. A., 15
Beerman, W., 251, 252
Belgium, 305
Bell Telephone Co and radio astronomy, 5, transistor research 11; computer, laser research 12
Ben Bella, 305
Berg, P. 155
Berry, J. P., 187
Berzelius, 3
Bikerman, J. J., 211, 212
Biology, humanistic 45-6, and man's nature and history 45-4, and man in the chain of being 48-52, and Paleolithic toolmakers

in man's nature, 53-58; and experiential innocence, 58-62; and capacity for self-creation, 62-66, 68
Black, Max, 220
Blum, H. F., 90
Bohr, Niels, 8, 9
Bolton, E. T., 164
Bonner, J., 252
Bose, R. C., 25
Boyland, E., 76, 77, 82, 84
Bradstreet, S. W., quoted, 175
Bragg, W. H., and W. L., 177
Brahe, Tycho, 374
Braithwaite, D. E., 192
Braithwaite, Richard, 220
Brancusi, quoted, 61
Brattain, W. H., 11
Brearley, W., 197
Brenner, S. S., 199
Bridgman, Percy, 11
Bristol University, 4
British American Oil Co., 19
British Gas Council, 286, 287, 298
British Gas Industry, 281
British Methane, Ltd., 287
Brittle materials, strengths and weaknesses, 171-216; frame of reference for, 172-74; brittleness and ductility, 174-75; ceramics and glass for study, 175-76; and structure of glass, 176; random network theory, 177-79; other theories of glass structure, 179-80; and estimates of theoretical strength, 180-83; Inglis and stress concentration, 183-84; and Griffith theory, 185-87; and observed strength of silicate glasses, 187-90; statistical theories, 190-92; and standardization of test conditions, 192-93; and static fatigue, 193-95; four ranges of glass strength, 196-99; and current status of the microcrack, 199-200; and ion exchange, 200-02; and pulsed stress experiments, 202-04; preliminary conclusions and new problems, 204; and calculation of Young's modulus, 204-06; data for glasses, 206; dependence of strength, 206-08; data for polycrystalline ceramics, 208-10; Inglis and Griffith criteria re-examined, 210-12; references, 213-16
Bronowski, J.: *The Identity of Man*, 217, 226, 232, quoted, 218-19, 233-34, 236; *The Common Sense of Science*, 222
Brooks, Harvey, quoted, 317
Brunauer-Emmett-Teller method, for defining surface area, 4
Bureau of Mines, U.S., 280, 282
Burns, R. M., 13
Bush, Brian B. H., 334
Bush, Vannevar, 12
Buu-Hoi, N. P., 89

Cabot, 280
Cady, H. P., 280
Cairns, J., 139, 141, 148, 157
Cambridge University, 4, 220
CAMEL (Cie Algierienne du Methane Liquid), 298
Canada, new research community planned, 18-19
Carnap, Rudolph, 222
Caro, Lucien, 137, 142
Catalysis, 2-4
Cattell, James McKeen, 106
Chalmers, P., quoted, 46
Chandra, G. R., 252
Charles, R. J., 172, 183, 187, 195, 205
Chechulin, B. B., 192
Chemical carcinogens, carcinogenesis, and carcinostasis, 69-103; progress in study of carcinogenesis, 69-74; theories of carcinogenesis, 74-76; indirect carcino-

INDEX

gens, 76-77, cancer-forming DNA, 77-79, carcinogenesis and energy transfer, 79-81, K-L-M region theory, 81-86, polycyclic compounds, study of, 86-87, tumor growth inhibition by, 88-90, photodynamic activity and carcinogenic action, 90-94, and synthesis of new polycyclic heterocyclic compounds, 95-98, references, 98-103
Chicago Bridge and Iron Co, 313
Church, Alonzo, 221, 222, 226
Clayson, D B, 75
Clements, J F., 208
Clever, U., 251
Color names, 105-32, reasons for studying 107-08, and dimensions of color space, 108 11, total number of discernible colors, 112 and absolute vs comparative judgments, 112-13, and differences in individual perception 113-14, number of, 114-16, in common use, 116; denotative meaning, 116-17, chosen for test, 117-18, testing procedure, 118, observers for test 118-19, basic form of data, 119-20, consistencies, 120, of selections by the two sexes, 122, for groups of names, 122-23, and mean hue selections, 123-26, overlap of, 126-27, total number of distinctive, 123-29; concluding observations on, 129-30; references 131-32
Conch International Methane Ltd 297, 298, 301, 306, 309, 318
Consolidated mass spectrograph, 14
Consolidated Mining and Smelting Co., 15
Comstock International Methane Ltd., 285, 288, 295-97, 305, 306, 309, 316
Comstock Liquid Methane Corp 286
Comstock-Prichard, Inc., 309 312, 318
Continental Oil Co, 285, 297, 315, 318
Cooper, H C., 281
Coporali, R. V., 206
Cornell University, 146
Corning Glassworks, 173, 187, 196, 212
Cowan, J D, 333
Cretan paradox, 227, 228, 231
Cromwell, N H, 93
Curie, Marie and Pierre, 7

Daniell, J F 160
Dao, T L 89
Davis, W R, 208
Davisson J W 9
Davy Edmund 3
Davy Humphrey 3
De Boer, J cell theory of liquid helium 30-37
De Broglie Louis 9 22 37 38
Debye theory 27 31 32 36 37 38 39, 40 41 42 48
Department of Agriculture U S 4
Descartes Rene 47 228 229 232
Dodds The Greeks and the Irrational 57
Donach I 90
Dresser Industries Ltd 282
Dunlop International Research 18-19
DuPont Co 13

East Ohio Gas Co 281-82
Ehrlich, G quoted 15-16
Einstein, Albert, 8 25 224
Elliot, H A 186
Emerson, Ralph Waldo quoted 60
England, 286 *See also* Great Britain
ENIAC, computer 12
Epstein, B, 190
Epstein S S, 91, 92 93 94
Ernsberger, F M 172 199 200-01 202, 212, quoted, 173
Evans, Pauline 116

Falk, H. L., 77
Faraday, M., 3
Fenstermacher, J. E., 208
Fermi, Enrico, 7, 8, 25
Feynman, R. P., interpretation of superfluidity, 29-30
Field, J. E., 203
Fisher, J. C., 172
Floridin Co., 278
Forro, F., 142
France, 286, 299, 305
Freud, paradoxes in, 230-31
Frick Chemical Co., 14
Frisch, K. von, 357
Fuita, T., 86

Gajewski, W., 248
Galileo, 6
Gamble Brothers, Inc., 286, 297
Gene to character in higher plants, 239-71; role of enzymes, 240-47; flower form and function, 247-49; gene action and differentiation, 249-53; differentiation at single-cell level, 253-56; genic control of cell-division patterns, 256-69; references, 269-71
General Electric Co., 4, 11, 15
Geneva Conference on the Peaceful Use of the Atom (*1955*), 273
Germany, 286, 305
Germer, L. H., 9, 16
Gilman, J. J., 181
Gödel, Kurt, theorems, 220-21, 222, 225, 226, 227, 229, 235, 236
Goldsmith, T. H., 356
Gordon, J. E., 199
Great Britain, 286, 298, 299, 303
Green, B., 75, 82
Green, H. N., 88
Greene, C. H., 173; quoted, 192
Griffith, A. A., 176; theory of stress concentration, 185-87, 299, 207, 210-12
Grodins, F. S., 360
Gurney, C., 186

Hansch, C., 86
Harvard University, 12
Hasselman, D. P. H., 209
Hassenstein, B., 360
Hassi R'Mel gas field, 298
Heidelberger, C., 84
Helium, liquid, 21-43; quantum aspects, 21-22, 24; heat conductivity, 22-23; viscosity, 23-24; isotopes, 24-25; and Landau theory, 28, 32, 36; Feynman interpretation, 29-30; cell theories, 30-37; extension of Debye theory proposed, 37-39; references, 43
Helmholtz, H. von, 332
Hilbert, David, *Entscheidungsproblem,* 220, 221
Hillig, W. B., 172, 182, 186, 195, 198, 204
Hobbes, Thomas, 230
Hoffman, D., 88
Holley, R. W., 146
Holloway, D. G., 197
Holmdel, New Jersey, 5
Holst, E. von, 344
Hope Natural Gas Co., 281
Horizons Inc., 212
Hoyer, B. H., 164
Huang, R. C., 252
Huggins, C., 82
Hulett, G. A., 13
Hume, David, 230
Huntington, R. L., *quoted,* 283

Industry. *See* Academia and industry
Ingall's Ship Yard, 285
Inglis, C. E., studies in stress concentration, 183-84, 186, 210
Institute for Advanced Studies, 12
International Lead Zinc Research Organization, 212
International Nickel Company of Canada, 18
Inter-Society Color Council, 124
Italy, 286

Methane, properties of, 278-80
Methane Pioneer (tanker), 288, 294, 295, 296, 301, 302
Methane Princess (tanker), 301, 304
Methane Progress (tanker), 301, 304
Mexico, 317
Miller, E. C., 76, 85
Miller, J. A., 76, 85
Misener, A. D., 18
Mittelstaedt, H., 353
Monod, J., 245
Moody, M. F., 356
Moon, 6
Moriconi, E. J., 87
Morley, J. G., 198
Morrison, Willard, 284
Mossbauer effect, 11
Mottram, J. C., 90
Mould, R. E., 187, 191, 206, 212; observations on static fatigue, 193, 195
Munier, J. H., 212
Munsell *Book of Color,* 118
Murphy, J. A., 285, 286

Nakayama, J., 199
NAM (Nederlandse Aardolie Maatschapij), 304
Naray-Szabo, I., 182
Nash, T., 79, 80
National Academy, Space Science Board, 377
National Aeronautics and Space Administration, U.S., 313, 314; selection of space experiments, 376-82
National Bureau of Standards, U.S., 111, 114, 117, 124
National Institutes of Health, 78
Netherlands: natural-gas field discovered, 275-76, 304-05; pipline expense, 306-07
Neumann, John von, *The Computer and the Brain,* 12
Neville, Harvey, 13
New Scientist, 16
Nietzsche, Friedrich, *The Birth of Tragedy,* 57
Nuclear science, academia and industry in, 6-8

Ohio Oil Co., 306
Ontario, Canada, 19
Ontario Research Foundation, 18
Orowan, E., 172, 186
Ortega y Gasset, José, quoted, 45
Otto, W. H., 191, 192

Parker, C. J., 212
Parriss, J. R., 356
Pennsylvania State College, 152, 157
People's Gas Co., 283
Phillips, C. J., 173, 193, 205, 206, 208
Phillips Petroleum Co., 309
Physical Review, 12
Pittsburgh Plate Glass Co., 212
Plato, *Phaedrus,* 57
Plochere Color System, 115
Poincaré, Henri, 222
Polar LMG Corp., 306
Poncelet, E. F., 186, 211
Popper, Karl, 222, 230-31
Preston, F. W., 193, 196, 212; quoted, 192
Prince, William Wood, 284, 285, 296
Princeton University, 4, 13
Procter, William, 1
Prod'homme, L., 180
Provance, J. D., 199
Pullman, A. and B., 80, 84, 87

Quantum mechanics, 9; and city traffic, 25-26; and musical instruments, 26-27; Landau superfluid theory, 28
Quastler, H.: *Information Theory in Biology,* 147; Principle of Signatures, 330-32

Raab, O., 90
Rabi, I. I., quoted, 16

Radio astronomy, 5-6
Ramsey, Frank, 220, 225
Ranger VII, 5
Rice, J M, 94
Richard, Jules, 227
Richter, M, 105
Rideal, E K, 12
Royal Dutch/Shell, 297, *304*, 305
Royal Society of London, expenditures, grants, 17-18
Russell, Bertrand, 227, 228
Rutherford, E, 7, 165, quoted, 166
Ryshkewitch, E, 199

Sabatier, 3, 4
San Diego Gas and Electric Co, 313
Saturn, 6
Schellinger, A K, 211
Schone, H, 344
Schroedinger, 9, 29, 30, 142
Schurkow, S, 173
Schurr, S H, quoted, 273
Schwalbe, W L, 206, 209
Schweet, R., 153
Scientific Research Society of America, 1-2
Selenium-Tellurium Development Association, Inc, 212
Shah, S S, 267
Shand, E B, 172, 187 189, 190, *211*, *212*
Shannon, C E, 330, 331
Sharp, H R, 311
Shaver, W W, 212
Sheridan Park, Ontario, 19
Schockley, W, 11
Sigma Xi, Society of, founds RESA, 1-2
Silverman, W B, 209
Sloan-Kettering Institute for Cancer Research, 77
Smoke, E J, 208
Smoluchowski, R, 152
Smyth, H D, quoted, 8
Smyth, H T, 212
South Africa, 306
Southwick, R D, 187, 193
Space experiments selection of, 373-82, and government role as patron, 373-76; and NASA program, 376-82, three classes of experiments, 377, expenses of, 378-79, and long lead time for experiments, 379-80, and public attention, *380-81*, and systems engineering concepts, 381-82, need for broader base of scientific participation, 382
Spinner, S, 205
Sporn, P, 273
Spriggs, R M, 191
Standard Oil of New Jersey, 304, 305
Stanford University, 18
Stephenson, J L, 342
Sucov, E W, 192
Sueoka, N, 148
Sun, 6, radio signals from, 5
Sung, Shou-Sin, 80
Sweden, 286, 305
Symmers, C, 197-98, 201
Symposium on Systems Analysis in Biology (*1966*), 363
Systems analysis and visual orientation of animals, 323-72, basic control system, 324-26, a steering control model, 326-28, and nature of the reference value, 329, and relevance of information theory, 330, Quastler's Principle of Signatures, 330 32, and reliability problem, 332-33, current research program, 333-37, and visual data processing, 338-42, and efferent nerve traffic, 343-44, and azimuth orientation 344-46, polarization analyzer, 346-54, model for *e*-vector orientation 354-59, and further work on orientation, 359-60, and more complex orientation, 360-61, multidisciplinary approach to, 361-62, addendum (post-*1964* progress), 362-65, reference, 365-72

Szent-Györgyi, A., 79

Tanaka, Y., 89
Tarski, Alfred, 221, 222, 223, 226, 229, 235
Taylor, Guy B., 13
Temperley, H. N. V., cell theory of liquid helium, 30-37
Textile Color Card Association, Standard Color Card, 115
Thomas, W. F., 191, 192, 197, 198
Thomson, G. P., 9
Thomson, J. J., 7
Tillet, J. P. A., 187
Tillich, Paul, quoted, 66
Toepler vacuum pump, 13
Transcontinental Gas Pipeline Corp. (Transco), 312
Tsien, L. C., 199
Turing, A. M., 221, 222, 226
Twomey, L., 281
Tyte, L. C., 174

Union of Soviet Socialist Republics: lasers and masers, 12; natural-gas liquefaction, 282-83
Union Oil and Marathon Oil Co., 306
University of Kansas, 280
University of Maryland, 12
University of Sydney, new radio telescope, 5-6
Urey, H. C., 11

Van Tubergen, 142
Varner, J. E., 252
Venezuela, 306, 317
Venus, radio signals from, 5

Walton, W. H., 172, 208
Warren, B. E., 177, 178
Watanabe, N., 204, 206
Waterman, T. H., 353
Waves and particles, academic and industrial research on, 8-9
Weir, J. A., 246
Wentworth, W. E., 81
Whitehead, Alfred North, 228
Wiersma, C. A. G., 334
Wigner, Eugene P., quoted, 63-64
Willis, R. A., 69
Wisconsin Natural Gas Co., 313
Wittgenstein, Ludwig, *Blue Book*, 232
Wordsworth, William, 231, 232, 233
World War II, effect on scientific progress, 373-74
Wynder, E. L., 88

Yagil, Ezra, 262
Yale University, 142
Yang, N. C., 82
Yoshikawa, H., 148

Zachariasen, W. H., 177, 178
Zimmerman, A. M., 267